Discovering Features with Synergistic Interactions in Multiple Views

Chohee Kim [1] **Mihaela van der Schaar** [2 3] **Changhee Lee** [1]

Abstract

Discovering features with *synergistic* interactions in multi-view data, that provide more information gain when considered together than when considered separately, is particularly valuable. This fosters a more comprehensive understanding of the target outcome from diverse perspectives (views). However, despite the increasing opportunities presented by multi-view data, surprisingly little attention has been paid to uncovering these crucial interactions. To address this gap, we formally define the problem of selecting synergistic and non-synergistic feature subsets in multi-view data, leveraging an information-theoretic concept known as *interaction information*. To this end, we introduce a novel deep learning-based feature selection method that identifies different interactions across multiple views, employing a Bernoulli relaxation technique to solve this intractable subset searching problem. Experiments on synthetic, semi-synthetic, and real-world multi-view datasets demonstrate that our model discovers relevant feature subsets with synergistic and non-synergistic interactions, achieving remarkable similarity to the ground truth. Furthermore, we corroborate the discovered features with supporting medical and scientific literature, underscoring its utility in elucidating complex dependencies and interactions in multi-view data.

1. Introduction

Synergistic interaction offers information gain we attain from a relationship when attributes are considered *together*, which cannot be achieved when they are examined *separately* (McGill, 1954). A straightforward example is the exclusive OR (XOR) operation between two variables, where one must understand the values of both variables to accurately identify the outcome. In other words, knowledge of an individual variable alone is insufficient to provide adequate information about the target outcome of interest.

Unraveling features with synergistic interactions is particularly valuable in understanding complicated mechanisms comprised of multi-view data, where each view represents a set of different features, potentially from diverse modalities or sources, describing the same object. A typical example is deciphering cellular function mechanisms from multi-omics data (Bunnik & Le Roch, 2013; Vidal et al., 2011). For instance, diseases rarely arise from the malfunction of a single gene product. Instead, they often stem from the complex interactions of thousands of gene products (Barabási, 2007). Furthermore, the genes associated with a disease may individually have minor effects, but when combined, their cumulative impact can be significant (Petretto et al., 2007). These intricate interactions lead to diverse responses to treatment, underscoring the necessity for a comprehensive analysis of the interplay across multiple omic layers (Benson, 2016). For example, heat shock proteins are frequently deregulated in breast cancer, but their implications vary depending on the cellular context: specific microRNAs are up-regulated in cases of poor prognosis, whereas certain genes are co-expressed in instances of favorable prognosis (Buttacavoli et al., 2021).

Furthermore, considering that fewer than 1% of gene products are involved in such pathological pathways, pinpointing the responsible genomes and their interplay is essential for comprehending the underlying mechanism. This understanding, in turn, facilitates effective clinical decision-making tailored to individual patients (Jackson et al., 2018). Thus, while multi-view data presents many opportunities, it also requires careful and precise analysis, considering relationships between features across different views to fully leverage its potential.

Unfortunately, existing multi-view feature selection methods fail to capture the essence of crucial interactions across diverse views. Their primary focus lies on simplistic approaches such as concatenating multiple views and applying traditional single-view feature selection methods (Acharya et al., 2020; Zhang et al., 2019), or capturing limited multi-view interactions based on pre-defined knowledge graphs

[1]Chung-Ang University, South Korea [2]University of Cambridge, UK [3]The Alan Turing Institute, UK. Correspondence to: Changhee Lee <chl8856@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

(Pfeifer et al., 2022; Acharya et al., 2020) or highly correlated feature subsets (Lindenbaum et al., 2021). These limitations hinder our ability to uncover meaningful interactions across views that, when considered together, offer valuable insights into the target outcome.

Contributions. We propose a novel deep learning approach for multi-view feature selection capable of uncovering informative features by decomposing them into components reflecting *synergistic* and *non-synergistic* interactions across different views. Our method adopts an information-theoretic concept, called *interaction information*, to define a novel multi-view synergistic and non-synergistic feature selection problem. Throughout the experiments on synthetic, semi-synthetic, and real-world multi-view datasets, we validate that our proposed method discovers relevant feature subsets with different interactions across multiple views, outperforming traditional state-of-the-art single-view feature selection methods. Additionally, we demonstrate the efficacy of our method by introducing a variant of the Shapely value (Lundberg & Lee, 2017) to quantify the synergistic attribution of selected feature subsets when the ground truth interactions are unknown. Then, we further corroborate the discovered synergistic features with supporting medical and scientific literature.

2. Related Work

Feature Selection. Feature selection is a well-established area of study, largely categorized into three methods based on the search strategy: filter (Kira & Rendell, 1992; Liu et al., 1996), wrapper (Kohavi & John, 1997), and embedded methods (Tibshirani, 1996). Recently, several deep learning-based embedded feature selection methods have been proposed. These approaches leverage the capability of deep neural networks to capture more heterogeneous and complex interactions between input features and the target outcome. These methods learn to solve the non-differentiable subset selection problem by approximating it with Lasso penalization (Li et al., 2016), or, more recently, with continuous relaxation employing Bernoulli distribution such as Hard-Concrete (Lee et al., 2021; Imrie et al., 2022) and Gaussian (Yamada et al., 2020; Yang et al., 2022) distributions.

Multi-View Feature Selection. Despite advancements in feature selection methods, multi-view feature selection remains relatively under-explored. Most studies concatenate multi-view data into a single view and apply traditional (single-view) feature selection methods, focusing on identifying the most essential factors across different views (Acharya et al., 2020; Zhang et al., 2019). However, these methods often overlook the intricate interactions that emerge when multiple views are present together, which are not observable when each view is considered separately or merely concatenated into one single view.

A few studies have addressed feature selection in multi-view data without resorting to simple concatenation. These approaches utilize various techniques to integrate multi-view data, including discovering crucial features through knowledge graphs built upon feature relationships (Pfeifer et al., 2022; Acharya et al., 2020), selecting highly correlated feature subsets between two views using canonical correlation analysis (CCA) (Lindenbaum et al., 2021), and identifying significant features or embeddings (Yang et al., 2023; Lin & Zhang, 2023) by considering both shared and distinct information from multi-view data. While these methods offer insight into specific feature dependencies, they have limitations in capturing a more comprehensive understanding of how features interact across multiple views and contribute to the target outcome.

Feature Selection with Group Interactions. "CompFS (Imrie et al., 2022) is a single-view feature selection method that emphasizes the significance of interactions. It identifies groups of informative features, known as composite features, by selecting groups where the predictive power significantly diminishes upon removing any single feature within that group. While CompFS focuses on single-view data and does not directly analyze feature interactions across multiple views, it is the closest comparison to our work due to the limited existing research on discovering feature interactions across multiple views.

3. Preliminaries: Information-Theoretic View

3.1. Embedded Feature Selection

Let $\mathbf{X} \in \mathbb{R}^p$ and $Y \in \mathcal{Y}$ be random variables representing p-dimensional input features and the corresponding target label, respectively, whose realizations are denoted as $\mathbf{x} = (x_1, \ldots, x_p)$ and y, i.e., $(\mathbf{x}, y) \sim p_{XY}$. Here, $\mathcal{Y}$ is the label space where the target task becomes a regression task when $\mathcal{Y} = \mathbb{R}$ or a C-class classification task when $\mathcal{Y} = [C]$.

Embedded feature selection aims to select a minimal subset, i.e., $\mathcal{S} \subseteq [p]$, of features that are informative about the target label, which can be formally quantified as the mutual information (MI) between the selected feature subset and the target. Denote $\mathbf{g} = (g_1, \ldots, g_p) \in \{0,1\}^p$ to be a binary gate vector where g_d indicates whether the d-th feature is selected in $\mathcal{S}$ or not, i.e., $g_d = 1$ if $d \in \mathcal{S}$ and $g_d = 0$ otherwise. Then, we can define the selected feature subset, $\mathbf{X}_S \in (\mathbb{R} \cup \{*\})^p$, as $\mathbf{X}_S = \mathbf{g} \odot \mathbf{X} + (1 - \mathbf{g}) \odot *$, where $*$ be any point not in $\mathbb{R}$ and $\odot$ is an element-wise multiplication.[1] Then, selecting the smallest informative feature subset can be achieved by solving the following optimization problem:

$$\underset{\mathbf{g} \in \{0,1\}^p}{\text{minimize}} \ \|\mathbf{g}\|_0 \quad \text{subject to } I_\theta(Y; \mathbf{X}_S) > \delta \qquad (1)$$

[1] We substitute $*$ with the mean of each feature, i.e., $* = \mathbb{E}[x_d]$ in this paper.

where I_θ indicates the estimated MI parameterized by θ which is optimized concurrently during the feature selection process, and δ constrains the minimum amount of information present in $\mathbf{X}_S$ about Y.

3.2. Synergistic and Non-synergistic Interaction

Interaction information is a generalization of the MI for more than two random variables (McGill, 1954), which measures the influence of a variable on the amount of information shared between the other variables. More specifically, given three random variables X, Y, and Z, the interaction information can be given as $I(Y;X|Z) - I(Y;X)$ where $I(Y;X)$ is the MI between X and Y and $I(Y;X|Z)$ is the conditional MI between X and Y given Z. Having the two MI quantities not being equivalent indicates the *presence* of interactions between X and Z regarding Y. Here, we will discuss two types of interactions, namely, the *synergistic* and *non-synergistic* interactions.

We define *synergistic interactions* between X and Z regarding Y as the positive interaction information, as follows:

Definition 3.1. (Synergistic Interaction) Variables X and Z have synergistic interaction with respect to Y when the two variables satisfy $I(Y;X|Z) > I(Y;X) \Leftrightarrow I(Y;X,Z) > I(Y;X) + I(Y;Z)$, which represents that X possesses more information about Y given Z than it does unconditionally.

Here, the equivalent inequality term is achieved by applying the chain rule for MI, suggesting that we can obtain more information about Y considering both X and Z *together* than considering X and Z *individually*.

Conversely, we define *non-synergistic interaction* between X and Z regarding Y as zero or negative interaction information, as follows:

Definition 3.2. (Non-Synergistic Interaction) Variables X and Z have non-synergistic interaction with respect to Y when the two variables satisfy $I(Y;X|Z) \leq I(Y;X) \Leftrightarrow I(Y;X,Z) \leq I(Y;X) + I(Y;Z)$, which implies that the information between X and Y becomes redundant given Z.

Thus, zero interaction information indicates that Z does not provide any additional information about the MI between X and Y. The negative interaction information, on the other hand, implies that the MI of X and Y rather decreases when conditioned on Z, suggesting that some information of $I(Y;X)$ can also be obtained from $I(Y;Z)$ alone.

4. Multi-View Synergistic Feature Selection

Our goal is to discover *synergistic* and *non-synergistic* relationships from *informative* features to enable a more comprehensive analysis for multi-view feature selection while maintaining superior predictive performance.

Notation. Let $\bar{\mathbf{X}} = (\mathbf{X}^1, \ldots, \mathbf{X}^V) \in \mathbb{R}^p$ be a random variable for the set of multi-view features from entire V views with $p = \sum_{v=1}^{V} p_v$ and $Y \in \mathcal{Y}$ for the target label. Here, $\mathbf{X}^v \in \mathbb{R}^{p_v}$ denotes a random variable for p_v-dimensional features from the v-th view whose realizations are denoted as $\mathbf{x}^v = (x_1^v, \cdots, x_{p_v}^v)$. We denote $\mathbf{X}^{\setminus v} = (\mathbf{X}^1, \cdots, \mathbf{X}^{v-1}, \mathbf{X}^{v+1}, \cdots, \mathbf{X}^V)$ as the complement features, comprising entire features other than those from the v-th view. Following the notations in Section 3.1, we denote $\mathbf{g}_S = (\mathbf{g}_S^1 \cdots \mathbf{g}_S^V)$ and $\mathbf{g}_N = (\mathbf{g}_N^1 \cdots \mathbf{g}_N^V)$ to be binary gate vectors, where $\mathbf{g}_S^v, \mathbf{g}_N^v \in \{0,1\}^{p_v}$ indicate synergistic and non-synergistic subsets from the v-th view, i.e., $\mathcal{S}^v, \mathcal{N}^v \subseteq [p_v]$, respectively. Then, the selected subset of features with synergistic interactions and that with non-synergistic interactions from the v-th view can be defined as $\mathbf{X}_S^v = \mathbf{g}_S^v \odot \mathbf{X}^v + (1 - \mathbf{g}_S^v) \odot *$ and $\mathbf{X}_N^v = \mathbf{g}_N^v \odot \mathbf{X}^v + (1 - \mathbf{g}_N^v) \odot *$, respectively. Similarly, we can denote $\bar{\mathbf{X}}_S$ and $\bar{\mathbf{X}}_N$ be the selected synergistic and non-synergistic feature subsets from the entire views, and $\mathbf{X}_S^{\setminus v}$ and $\mathbf{X}_N^{\setminus v}$ be the complement of the v-th view comprising the selected feature subsets from other views, respectively.

4.1. Synergistic Feature Selection

Building upon Definition 3.1, we formally define the multi-view features with synergistic interactions as follows:

Definition 4.1. (Multi-View Synergistic Features.) Let $\mathcal{S}^v$ be a subset of features for $v \in [V]$. Then, $\mathbf{X}_S^v$ and $\mathbf{X}_S^{\setminus v}$ have synergistic interactions with respect to Y when the following inequality is satisfied:

$$I(Y; \mathbf{X}_S^v | \mathbf{X}_S^{\setminus v}) > I(Y; \mathbf{X}_S^v). \tag{2}$$

Our goal is to find the minimum subsets of features, i.e., $\mathcal{S} = (\mathcal{S}^1, \ldots, \mathcal{S}^V)$, from V views that contribute to multi-view synergistic interactions, i.e., $I(Y; \mathbf{X}_S^v | \mathbf{X}_S^{\setminus v}) > I(Y; \mathbf{X}_S^v)$ for $v \in [V]$. Here, we can simplify the multiple-constrained problem, each constraint defined for each view, into a single-constrained problem based on the following proposition.

Proposition 4.2. *Let $\mathcal{S}^* = (\mathcal{S}^{*1}, \cdots, \mathcal{S}^{*V})$ be the minimum subset of features that satisfies the inequality below:*

$$I(Y; \bar{\mathbf{X}}_{S^*}) > \sum_{v=1}^{V} I(Y; \mathbf{X}_{S^*}^v). \tag{3}$$

Then, $\mathcal{S}^{*v}$ is also the minimum subset that satisfies (2), i.e., $I(Y; \mathbf{X}_{S^*}^v | \mathbf{X}_{S^*}^{\setminus v}) > I(Y; \mathbf{X}_{S^*}^v)$ for $v \in [V]$.

This asserts that selecting a minimum subset of features that satisfies (3) guarantees that the selected feature subset must contain multi-view synergistic interactions.

Finally, we can define the multi-view synergistic feature

selection problem as follows:

$$\begin{aligned} &\underset{\mathbf{g}_S^1,\dots,\mathbf{g}_S^V}{\text{minimize}} \ \|\mathbf{g}_S^1\|_0 + \cdots + \|\mathbf{g}_S^V\|_0 \\ &\text{subject to} \ \ I_\theta(Y;\bar{\mathbf{X}}_S) > \sum_{v=1}^{V} I_\theta(Y;\mathbf{X}_S^v), \end{aligned} \tag{4}$$

where I_θ indicates the estimated MI parameterized by θ. Here, (4) effectively generalizes our synergistic feature selection problem into multiple views by introducing a single constraint based on (i) the MI of Y with the entire multi-view features and (ii) that with the sum of marginal features.

4.2. Non-Synergistic Feature Selection

Similarly, we can define the multi-view *non-synergistic* feature selection problem as selecting minimum subsets of features from V views, aiming to achieve zero or negative interaction information, as follows:

$$\begin{aligned} &\underset{\mathbf{g}_N^1,\dots,\mathbf{g}_N^V}{\text{minimize}} \ \|\mathbf{g}_N^1\|_0 + \cdots + \|\mathbf{g}_N^V\|_0 \\ &\text{subject to} \ \ I_\theta(Y;\bar{\mathbf{X}}_N) \leq \sum_{v=1}^{V} I_\theta(Y;\mathbf{X}_N^v). \end{aligned} \tag{5}$$

Here, the selected non-synergistic features satisfy the reverse direction of the constraint in (4), indicating that the MI of Y with these features decreases when computed jointly, rather than when computed individually.

Challenges. Overall, our goal is to advance the traditional feature selection problem by not only identifying informative features but also unraveling the various multi-view interactions among these features. However, solving the objectives for the multi-view synergistic and non-synergistic feature selection problem presents the following challenges: First, searching over the discrete space for selecting features becomes intractable in the high-dimensional regime. Second, accurate estimation of the MI terms associated with our objectives is required while jointly selecting synergistic and non-synergistic features. Third, it is essential to ensure that the selected synergistic and non-synergistic features are mutually exclusive while being informative about the target.

5. Method

To address the challenges associated with selecting informative features with *synergistic* and *non-synergistic* interactions, we propose a multi-view embedded feature selection method based on interaction information, called Multi-View **Syn**ergistic **F**eature **S**elector (SynFS).[2] Our method comprises three key components as depicted in Figure 1:

- a set of V *view-specific synergistic selectors*, parameterized by $\boldsymbol{\mu}_S = (\boldsymbol{\mu}_S^1, \cdots, \boldsymbol{\mu}_S^V)$, each of which governs the selection of synergistic feature subset,
- a set of V *view-specific non-synergistic selectors*, parameterized by $\boldsymbol{\mu}_N = (\boldsymbol{\mu}_N^1, \cdots, \boldsymbol{\mu}_N^V)$, each of which governs the selection of non-synergistic feature subset,
- a set of *predictors* for synergistic (f_{ϕ_S}), non-synergistic (f_{ϕ_N}), and union (f_{ϕ_A}) feature subsets, each of which takes selected feature subsets as input and outputs the corresponding conditional distribution of Y.

5.1. Synergistic Selectors

The set of synergistic selectors aims to identify the subset of features with synergistic interactions across multiple views, as formulated in (4).

Continuous Relaxation. To transform the intractable combinatorial search problem in (4) into a search over binary random variables, we employ the stochastic relaxation using Gaussian distribution (Yamada et al., 2020). Specifically, we assume that the binary gate vectors introduced in Section 4 follow the Bernoulli distribution, which we approximate using a Gaussian-based continuous relaxation. Given $\boldsymbol{\mu}_S^v = (\mu_{S,1}^v, \dots, \mu_{S,p_v}^v)$, we define the synergistic gate vectors, $\mathbf{g}_S^v$, as the realization of the *relaxed* Bernoulli variable, where $g_{S,d}^v = \max(0, \min(1, \mu_{S,d}^v + \epsilon))$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then, we can convert the constrained combinatorial searching problem into an unconstrained continuous optimization introducing a Lagrangian multiplier, as follows:

$$\begin{aligned} \mathcal{L}_{sel}^{\boldsymbol{\mu}_S} = \mathbb{E}_{\mathbf{x},y}\Big[\mathbb{E}_{\mathbf{g}_S}\Big[\ell(y, f_{\phi_S}(\bar{\mathbf{x}}_{\mathbf{g}_S})) - \sum_{v=1}^{V} \ell(y, f_{\phi_S}(\mathbf{x}_{\mathbf{g}_S}^v)) \\ + \lambda_S \|\mathbf{g}_S\|_0\Big]\Big], \end{aligned} \tag{6}$$

where λ_S is a hyper-parameter that controls the selection sparsity (i.e., the larger λ_S the fewer features are selected). Here, $\mathbb{E}_{\mathbf{g}_S}[\|\mathbf{g}_S\|_0] = \sum_v^V \sum_{d=1}^{p_v} \mathbb{P}(g_{S,d}^v > 0)$ is the sum of activated gates which is equivalent to solving $\sum_v^V \sum_{d=1}^{p_v} \Phi(\frac{\mu_{S,d}^v}{\sigma})$ where Φ is the standard Gaussian CDF. We convert maximizing (minimizing) the MI term into minimizing (maximizing) the conditional entropy, as $I(Y;X_S) = H(Y) - H(Y|X_S)$ and $H(Y)$ is irrelevant to our optimization target. We denote the cross-entropy term as $\ell(y,\hat{y}) = \sum_c y_c \log \hat{y}_c$ for a C-way classification if $\mathcal{Y} = \{1, \cdots, C\}$ and $\ell(y,\hat{y}) = \|y - \hat{y}\|_2^2$ for a regression task if $\mathcal{Y} = \mathbb{R}$.[3] Please see Appendix A.2 for detailed derivation. It is worth highlighting that our objective (6) is differentiable with respect to $\boldsymbol{\mu}_S$.

5.2. Non-Synergistic Selectors

Similarly, we suppose that the non-synergistic gate vectors, $\mathbf{g}_N^v$, follow the relaxed Bernoulli distribution (Yamada

[2] https://github.com/choheeK/SynFS.

[3] For $\mathcal{Y} = \{1, \dots, C\}$, we will occasionally abuse notation and write y_c to denote the c-th element of the one-hot encoding of y.

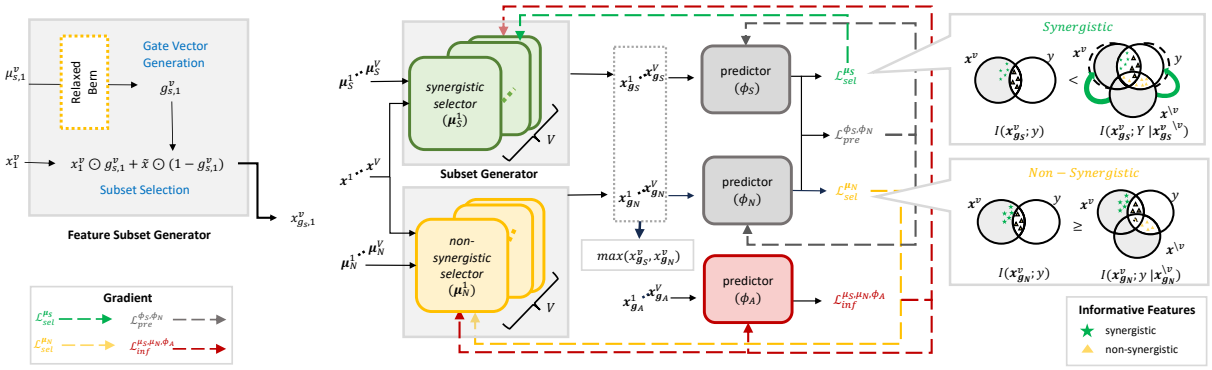


Figure 1: An overview of SynFS architecture. The two sets of view-specific selectors and three predictors are updated iteratively according to the corresponding loss and gradient highlighted in different colors.

et al., 2020) parameterized by $\boldsymbol{\mu}_N = (\boldsymbol{\mu}_N^1, \cdots, \boldsymbol{\mu}_N^V)$, where $g_{N,d}^v = \max(0, \min(1, \mu_{N,d}^v + \epsilon))$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Before transforming our non-synergistic feature selection objective in (5), we introduce an auxiliary loss function that encourages the selected synergistic and non-synergistic feature subsets to be mutually exclusive.

Mutual Exclusivity. Based on the definition of synergistic and non-synergistic feature subsets, a feature cannot simultaneously have both positive and negative interaction information with respect to Y. Thus, we introduce the following loss based on the similarity between the synergistic selector, $\boldsymbol{\mu}_S$, and the non-synergistic selector, $\boldsymbol{\mu}_N$ to ensure that the selected feature subsets from the two types of selectors are mutually exclusive: $\text{sim}(\mathbf{g}_S, \mathbf{g}_N) = \frac{\mathbf{g}_S \cdot \mathbf{g}_N}{\|\mathbf{g}_S\| \|\mathbf{g}_N\|}$.

Overall, using a Lagrangian approximation, we can convert (5) into the following objective:

$$\mathcal{L}_{sel}^{\boldsymbol{\mu}_N} = \mathbb{E}_{\mathbf{x},y}\Big[\mathbb{E}_{\mathbf{g}_N}\Big[\sum_{v=1}^{V} \ell(y, f_{\phi_N}(\mathbf{x}_{\mathbf{g}_N}^v)) - \ell(y, f_{\phi_N}(\bar{\mathbf{x}}_{\mathbf{g}_N})) + \lambda_N \|\mathbf{g}_N\|_0 + \alpha \cdot \text{sim}(\mathbf{g}_S, \mathbf{g}_N)\Big]\Big], \tag{7}$$

where λ_N is a coefficient that controls selection sparsity and α is a hyperparameter introduced to enforce mutual exclusivity between the two selectors. Note that we include the similarity loss only in (7) because having both synergistic and non-synergistic gates open for the same feature implies that this feature is more informative about the target when considered together with other views, aligning with the definition of a synergistic relationship.

5.3. Toward Informative Selected Feature Subsets

Now, we encourage the selected synergistic and non-synergistic feature subsets to be informative about the target. To this goal, we introduce a new gate vector, $\mathbf{g}_A$, whose element is the maximum of the corresponding elements in the synergistic and non-synergistic gate vectors, i.e., $g_{A,d}^v = \max(g_{S,d}^v, g_{N,d}^v)$. This naturally provides the union set of the two selected feature subsets represented by the relaxed gate vectors. Hence, we focus on increasing the MI between the union feature subsets and the target by maximizing the following objective:

$$\mathcal{L}_{inf}^{\boldsymbol{\mu}_S, \boldsymbol{\mu}_N, \phi_A} = \mathbb{E}_{\mathbf{x},y}\big[\mathbb{E}_{\mathbf{g}_S, \mathbf{g}_N}\big[\ell(y, f_{\phi_A}(\bar{\mathbf{x}}_{\mathbf{g}_A}))\big]\big]. \tag{8}$$

5.4. Predictors

As mentioned earlier, our goal is to choose feature subsets that minimize (maximize) cross-entropy terms, serving as a proxy for maximizing (minimizing) the MI terms of our interest. This requires accurate estimates of the cross-entropy terms, which can be obtained by correctly predicting the target based on the corresponding synergistic, non-synergistic, and union feature subsets. To achieve this, we focus on optimizing the predictors for the synergistic feature subsets (f_{ϕ_S}), for the non-synergistic feature subsets (f_{ϕ_N}), and for the union feature subsets (f_{ϕ_A}) based on the following objectives, respectively:

$$\mathcal{L}_{pre}^{\phi_S, \phi_N} = \mathbb{E}_{\mathbf{x},y}\Big[\mathbb{E}_{\mathbf{g}_S, \mathbf{g}_N}\Big[\ell(y, f_{\phi_S}(\bar{\mathbf{x}}_{g_S})) + \sum_{v=1}^{V} \ell(y, f_{\phi_S}(\mathbf{x}_{\mathbf{g}_S}^v)) + \ell(y, f_{\phi_N}(\bar{\mathbf{x}}_{\mathbf{g}_N})) + \sum_{v=1}^{V} \ell(y, f_{\phi_N}(\mathbf{x}_{\mathbf{g}_N}^v))\Big]\Big]. \tag{9}$$

Here, we have omitted the objective for the predictor, f_{ϕ_A}, as it is already described in (8). Please refer to Appendix A.2 for detailed derivations.

5.5. Training

Overall, we optimize the three components of SynFS, incorporating the previously defined objectives, given as follows:

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}_{sel}^{\boldsymbol{\mu}_S} + \mathcal{L}_{sel}^{\boldsymbol{\mu}_N}}_{\text{selector}} + \overbrace{\mathcal{L}_{inf}^{\boldsymbol{\mu}_S, \boldsymbol{\mu}_N, \phi_A}}^{\text{predictor}} + \mathcal{L}_{pre}^{\phi_S, \phi_N}. \tag{10}$$

Table 1: Optimization directions for different objectives.

	Joint views	Marginal views
Predictor (ϕ_N, ϕ_S)	$\ell(y, f_{\phi_S}(\bar{\mathbf{x}}_{\mathbf{g}_S}))\downarrow$ $\ell(y, f_{\phi_N}(\bar{\mathbf{x}}_{\mathbf{g}_N}))\downarrow$	$\sum_v \ell(y, f_{\phi_S}(\mathbf{x}^v_{\mathbf{g}_S}))\downarrow$ $\sum_v \ell(y, f_{\phi_N}(\mathbf{x}^v_{\mathbf{g}_N}))\downarrow$
Synergistic selector $(\boldsymbol{\mu}_S)$	$\ell(y, f_{\phi_S}(\bar{\mathbf{x}}_{\mathbf{g}_S}))\downarrow$	$\sum_v \ell(y, f_{\phi_S}(\mathbf{x}^v_{\mathbf{g}_S}))\uparrow$
Non-synergistic selector $(\boldsymbol{\mu}_N)$	$\ell(y, f_{\phi_N}(\bar{\mathbf{x}}_{\mathbf{g}_N}))\uparrow$	$\sum_v \ell(y, f_{\phi_N}(\mathbf{x}^v_{\mathbf{g}_N}))\downarrow$

Here, accurately estimating the MI terms while jointly selecting synergistic and non-synergistic feature subsets is challenging, as some parts of the objective have conflicting directions for their optimizations. More specifically, the synergistic selectors, $\boldsymbol{\mu}_S$, are designed to choose more predictive features that offer additional information about the target when considered jointly across all views. This combined information should exceed what can be gained by analyzing the features individually Optimizing synergistic selectors involves minimizing the cross-entropy of the target variable given feature subsets from the joint views, while simultaneously maximizing the cross-entropy for the sum of marginals. In contrast, the corresponding predictor for synergistic feature subset, f_{ϕ_S}, minimizes the cross-entropy of the target given feature subsets from both joint and marginal views. Such conflicting optimization direction can make synergistic selectors fall into a trivial solution by adversely impairing the estimation of marginal views. We summarize the optimization direction for each component in Table 1.

To avoid such issues, we ensure that parameters are updated *iteratively* based on their corresponding objectives as depicted in Figure 1. Please refer to Appendix A.3 for the overall training procedure of SynFS.

6. Experiments

In this section, we evaluate the performance of SynFS and multiple feature selection methods using synthetic, semi-synthetic, and real-world experiments. The primary goal of these experiments is to demonstrate the capability of our method in discovering features with multi-view interactions.

Benchmarks. To the best of our knowledge, there is no existing work on selecting feature subsets with multi-view interactions. Hence, we have compared SynFS with the state-of-the-art feature selection method, called **CompFS** (Imrie et al., 2022), which aims to select feature groups, each comprising features with interactions, in a single-view setting. We also compare our method with **STG** (Yamada et al., 2020) and Random Forest (**RForest**) (Ho, 1995), which are deep learning-based and ensemble-based methods commonly used for traditional feature selection, respectively. Please see Appendix A.9 for more details.

Performance Metrics: Known Ground Truths. When ground truth synergistic and non-synergistic features are known, we evaluate the discovered feature subsets of ours and CompFS using *Jaccard Index* (J-Index) (Jaccard, 1912). The J-Index of the two sets $\mathcal{A}$ and $\mathcal{B}$ can be given as $J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}$. Here, we report $\frac{1}{2}(J(\mathcal{S}, \hat{\mathcal{S}}) + J(\mathcal{N}, \hat{\mathcal{N}}))$ to evaluate the discovered synergistic and non-synergistic subsets, denoted as $\hat{\mathcal{S}}$ and $\hat{\mathcal{N}}$, respectively. For CompFS, we re-categorize interactions to synergistic when more than one features are selected from different views in one subset and non-synergistic otherwise. For the traditional feature selection methods, we apply the Group Similarity score as suggested in (Imrie et al., 2022), which computes a normalized J-Index by using the most similar subsets with ground truth. We also assess the true positive rate (TPR) and false discovery rate (FDR) of selected features to determine whether these features are important (either synergistic or non-synergistic, i.e., $\mathcal{S} \cup \mathcal{N}$). Please refer to Appendix A.10 for more details about the above performance metrics.

Performance Metrics: Unknown Ground Truths. Evaluating the performance of feature and interaction discovery on real data is difficult since ground truth relevance is rarely known. Therefore, we evaluate discovery performance based on two metrics. First, we assess the discriminative power of the selected feature subsets by training a separate MLP only utilizing the selected features and then comparing the AUROC. Second, We propose a *Set Interaction Score (SI)* to quantify the magnitude of interaction among selected features by assessing the difference in the AUROC performance when the features are considered collectively versus individually. We calculate the performance improvement based on all possible subsets of selected features using a Shapley-based method (Lundberg & Lee, 2017). That is, given a set of features $\mathcal{S} = \mathcal{S}^1 \cup \cdots \cup \mathcal{S}^V$, we define the SI as the average contribution of each feature in the set, as:

$$SI(\mathcal{S}) = \frac{C}{|\mathcal{S}|} \sum_{v \in V} \sum_{d \in \mathcal{S}^v} \sum_{a \subset \mathcal{S}^v \setminus \{d\}} P(a + \{d\}) - P(a) \quad (11)$$

where C is a normalizing factor $\frac{|a|!(|S^v|-|a|-1)!}{|S^v|!}$. Here, $P(a)$ for $a \subset \mathcal{S}^v$ approximates an interaction score and is defined as $P(a) = \mathtt{F}(\mathcal{S}) - \sum_{i \in V \setminus v} \mathtt{F}(\mathcal{S}^i) - \mathtt{F}(a)$, where $\mathtt{F}$ symbolizes a performance metric, such as AUROC, which is computed only by taking input subset. The intuition behind the SI is that a synergistic feature will have a larger positive value, as it will have a higher AUROC when considered *together* with features from other views than when considered *separately* in each view. This implies that a properly chosen synergistic feature set should have a larger SI. Please see Appendix A.11 for a detailed explanation of the proposed metric and its validation on the synthetic datasets.

6.1. Synthetic Experiments

Dataset Description. We start by evaluating the synergistic and non-synergistic feature selection performance by uti-

Table 2: Synthetic data generation process.

Dataset	Rule	Synergistic	Non-synergistic	Description
Syn1	$(\mathbf{X}^1[1] \cdot \mathbf{X}^2[2] > 0.3)$ OR $(\mathbf{X}^1[8] > 0.5)$ OR $(\mathbf{X}^2[8] > 0.5$	$\mathbf{X}^1[1], \mathbf{X}^2[2]$	$\mathbf{X}^1[8], \mathbf{X}^2[8]$	Synergistic and non-synergistic interactions in each view with a strong signal.
Syn2	$\mathbf{X}^1[1] * \mathbf{X}^2[2] + \mathbf{X}^1[3] + \mathbf{X}^2[4]$	$\mathbf{X}^1[1], \mathbf{X}^2[2]$	$\mathbf{X}^1[3], \mathbf{X}^2[4]$	Same interactions as in **Syn1**, but in a weaker signal where detecting interactions is more challenging.
Syn3	$\mathbf{X}^1[1] * \mathbf{X}^2[2] + \mathbf{X}^1[3] + \mathbf{X}^2[4]$	$\mathbf{X}^1[1], \mathbf{X}^2[2]$	$\mathbf{X}^1[3], \mathbf{X}^2[4]$	Same interactions as in **Syn2** in a 3-view setting, where View 3 is redundant.
Syn4	$(\mathbf{X}^1[1] + 2) \cdot \mathbf{X}^2[2] \cdot (\mathbf{X}^3[3] - 2)$	$\mathbf{X}^1[1], \mathbf{X}^2[2], \mathbf{X}^3[3]$	N/A	Complicated synergistic interactions across 3-views without any non-synergistic interactions.

lizing a set of synthetic datasets employed in (Chen et al., 2018) and (Imrie et al., 2022) with known ground truth. Specifically, the input features $\bar{\mathbf{X}} \in \mathbb{R}^{500}$ are generated from an independent Gaussian distribution, i.e., $\mathcal{N}(0, \mathbf{I})$, and the corresponding labels Y are generated based on the following two cases:

- Strong signal (Imrie et al., 2022): the label is derived based on the given decision rules.
- Weak signal (Chen et al., 2018): the label is sampled from a Bernoulli distribution where $\mathbb{P}(Y = 1|\bar{\mathbf{X}}) = \frac{1}{1+\texttt{logit}(\bar{\mathbf{X}})}$, with $\texttt{logit}(\bar{\mathbf{X}})$ generated by each rule.

The data generation rule and ground truth interactions for the different synthetic scenarios are summarized in Table 2. Multiplication of features from different views represents synergistic interaction as more information can be obtained by considering them together whereas linear summation represents non-synergistic interaction. We generate 20,000 samples with 500 features (250-250 for two views and 200-200-100 for three views). All the results are averaged over 10 random iterations of random 64/14/20 training/validation/testing splits.

Quantitative Analysis. Table 3 shows that SynFS consistently correctly discovers synergistic and non-synergistic interactions without any false discovery across all the tested synthetic scenarios, even when there is a redundant view without any important features (Syn3). STG demonstrates a perfect discovery of important features without any false discoveries (and nearly perfect for RForest). However, as traditional feature selection methods lack a mechanism to capture interactions between views, they result in poor normalized J-Index. Notably, CompFS almost perfectly distinguishes different interactions in Syn1 with a strong signal and performs relatively well in Syn4, where only synergistic features exist in each view. However, it exhibits limited performance in Syn2 and Syn3 when each view contains weak signals from both synergistic and non-synergistic features. Additionally, it often fails to select important synergistic features, leading to low TPR performance.

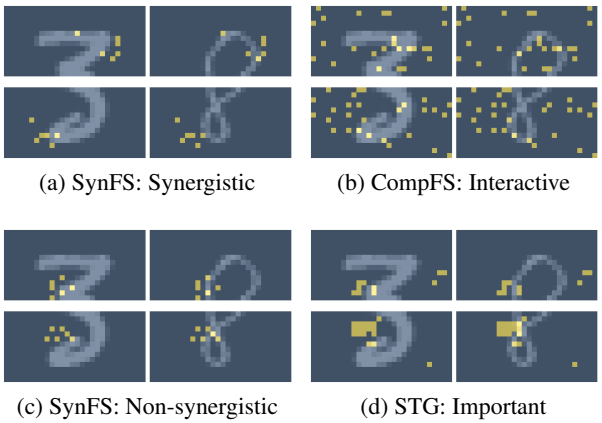
6.2. Semi-Synthetic Experiments: MNIST

Dataset Description. We illustrate the interaction discovery performance of SynFS on the MNIST (LeCun et al., 1998), using a commonly used feature selection scenario of distinguishing '3' and '8'. To simulate the multi-view setting, we horizontally divided each image into two parts, creating two views following the convention from previous works (Chen & Denoyer, 2017; Goyal et al., 2019).

Table 3: Performance results on the synthetic datasets. Here, † indicates the normalized J-Index.

Dataset	Methods	J-Index ↑	TPR ↑	FDR ↓
Syn1	SynFS	1.00 ± 0.00	100 ± 0.0	0 ± 0.0
	CompFS	0.91 ± 0.18	100 ± 0.0	0 ± 0.0
	STG	$0.47 \pm 0.07^{\dagger}$	100 ± 0.0	0 ± 0.0
	RForest	$0.50 \pm 0.00^{\dagger}$	100 ± 0.0	0 ± 0.0
Syn2	SynFS	1.00 ± 0.00	100 ± 0.0	0 ± 0.0
	CompFS	0.48 ± 0.25	70 ± 25.0	3.3 ± 10.0
	STG	$0.50 \pm 0.00^{\dagger}$	100 ± 0.0	0 ± 0.0
	RForest	$0.33 \pm 0.00^{\dagger}$	50 ± 0.0	0 ± 0.0
Syn3	SynFS	1.00 ± 0.00	100 ± 0.0	0 ± 0.0
	CompFS	0.40 ± 0.20	50 ± 0.0	0 ± 0.0
	STG	$0.50 \pm 0.00^{\dagger}$	100 ± 0.0	0 ± 0.0
	RForest	$0.50 \pm 0.00^{\dagger}$	50 ± 0.0	0 ± 0.0
Syn4	SynFS	1.00 ± 0.00	100 ± 0.0	0 ± 0.0
	CompFS	0.90 ± 0.20	100 ± 0.0	0 ± 0.0
	STG	$1.00 \pm 0.00^{\dagger}$	100 ± 0.0	0 ± 0.0
	RForest	$1.00 \pm 0.00^{\dagger}$	100 ± 0.0	0 ± 0.0

Qualitative Analysis. The synergistic features identified by SynFS in Figure 2a underscore our method's ability to detect subtle and previously unnoticed relationships by focusing on interactions in two views. More specifically, our method reveals that the diagonal pixels – particularly, the top-right and bottom-left regions – hold more information for distinguishing between '3' and '8', when considered together. Notably, these pixels seem to be less informative when considered without interaction information between the two views. On the other hand, the non-synergistic features discovered by SynFS are pixels in central regions that are chosen as most discriminative by STG, as shown in Figure 2d. This demonstrates that SynFS is capable of identifying the important features overlooked by previous methods and successfully distinguishing between two types of interactions. Figure 2b illustrates pixels discovered by CompFS, all of which are identified as synergistic without any non-synergistic features. However, it is challenging to visually understand how these specific pixels contribute to



(a) SynFS: Synergistic (b) CompFS: Interactive

(c) SynFS: Non-synergistic (d) STG: Important

Figure 2: An illustrative comparison of selected features.

the overall synergistic relationship.

We provide an additional set of experiments with artificially induced biases in Appendix A.6 to further validate whether our proposed method can correctly identify features with synergistic interactions.

6.3. Real-World Experiments: METABRIC

Dataset Description. METABRIC (Pereira et al., 2016) is a multi-view dataset comprising 1,980 primary breast cancer samples, where each sample is described by clinical data and targeted sequencing data including mRNA expression, copy number alteration, and mutation. We construct a two-view dataset utilizing mRNA expression ($p_1 = 489$) and mutation profiles ($p_2 = 177$), focusing on discovering crucial synergistic features between the two views that are informative about the progesterone receptor (PR) status.

Quantitative Analysis. As demonstrated in Table 5, SynFS outperforms traditional and interaction-oriented feature selection methods for interaction discovery performance metrics. Here, we treat all the features selected by STG and RForest as synergistic features and apply the same re-categorization for the features selected by CompFS. We leave $SI(\mathcal{N})$ blank when non-synergistic feature subsets are not available. Note that all the SI reported in Table 5 are calculated using 20 features from each view with the highest importance and are scaled by 100 for better readability. Notably, SynFS is the only method capable of accurately capturing view-synergistic features, providing positive $SI(\mathcal{S})$ value. Furthermore, non-synergistic features have large negative values, underscoring their contributions from an individual view perspective. In addition, SynFS achieved high predictive performance using only 6% of the entire features, with only a marginal performance gap compared to an MLP utilizing the entire features.

Qualitative Analysis. We provide supporting evidence from medical and scientific literature for the most frequently discovered synergistic features (more than 7 times out of 10 iterations) as summarized in Table 4. For instance, a synergistic interaction can be found between *KMT2C* mutation and *MAPT* gene expression. The *KMT2C* mutation can indicate up-regulation of *MAPK* signaling pathway (Wang et al., 2021; Zhao et al., 2011), which is also influenced by *MAPT* (Zhang & Liu, 2002). Notably, the *MAPK* pathway is known to activate PR, thereby playing a crucial role in the progestin-induced proliferation of breast cancer (Hagan et al., 2012).

Table 4: Supporting evidence for synergistic interactions discovered by SynFS on the METABRIC dataset.

View 1 (Gene Expression)	View 2 (Mutation)	Reference
MAPT	*KMT2C*	(Wang et al., 2021) (Zhao et al., 2011) (Zhang & Liu, 2002)
NRIP1	*KMT2C*	(Binato et al., 2021) (Al-Khayyat et al., 2023)
TGFBR3	*PDE4DIP*	(Liu et al., 2006) (Dodge et al., 2001) (Ogiwara et al., 2021)
HSD17B7	*PDE4DIP*	(Wang et al., 2015) (Aronica et al., 1994)
BIRC6	*KMT2C*	(Wu et al., 1993) (Li et al., 2021)

6.4. Real-World Experiments: TCGA

Dataset Description. We analyze 1-year mortality based on the comprehensive observations from two omics layers (i.e., microRNA and RPPA), considering each layer as a separate view, on 7,295 cancer cell lines collected by the Cancer Genome Atlas (TCGA)[4].

Quantitative Analysis. Table 5 compares the predictive power of feature subsets selected by SynFS and the baselines. For SynFS, we select features based on a threshold of 0.55 across all views, which maximizes $SI(\mathcal{S}) - SI(\mathcal{N})$ on the validation set, resulting in 154 features in total. For benchmarks, we select the same number of features in their order of importance to provide a fair comparison. SynFS demonstrates competitive predictive performance while being the only method capable of distinguishing different types of synergistic and non-synergistic interactions among views, showing the high $SI(\mathcal{S})$ and the low $SI(\mathcal{N})$.

Qualitative Analysis. Table 6 summarizes the supporting evidence for potential synergistic interactions between RPPA proteins and miRNAs in determining 1-year mortality in cancer patients. This evidence focuses on the most frequently identified synergistic features by *SynFS* after 10

[4]https://www.cancer.gov/tcga

Table 5: Performance results on the real-world datasets.

Dataset	Methods	AUROC ↑	$SI(\mathcal{S})$ ↑	$SI(\mathcal{N})$ ↓
METABRIC	SynFS	0.862 ± 0.02	0.36 ± 0.13	-3.25 ± 0.40
	CompFS	0.846 ± 0.03	-0.71 ± 1.08	0.00 ± 0.00
	STG	0.887 ± 0.01	$-0.22 \pm 0.46^{\dagger}$	-
	Rforest	0.859 ± 0.01	$-0.38 \pm 0.94^{\dagger}$	-
	MLP	0.884 ± 0.01	-	-
TCGA	SynFS	0.683 ± 0.03	7.87 ± 18.80	-19.65 ± 18.24
	CompFS	0.686 ± 0.03	-2.27 ± 0.59	0.00 ± 0.00
	STG	0.680 ± 0.01	$-1.65 \pm 0.35^{\dagger}$	-
	Rforest	0.677 ± 0.02	$-2.25 \pm 0.78^{\dagger}$	-
	MLP	0.697 ± 0.01	-	-
PBMC	SynFS	0.984 ± 0.00	0.38 ± 1.86	-460.2 ± 1.42
	CompFS	0.987 ± 0.00	-20.02 ± 0.71	0.00 ± 0.00
	STG	0.980 ± 0.00	$-21.41 \pm 0.72^{\dagger}$	-
	Rforest	0.984 ± 0.00	$-21.80 \pm 1.51^{\dagger}$	-
	MLP	0.977 ± 0.00	-	-

iterations. For example, *Caspase-7*, a key protein in apoptosis (programmed cell death) has been shown to interact with various miRNAs (Su et al., 2015). Specifically, *Caspase-7* can inactivate the *EGFR* signaling pathway, which regulates critical cellular events like proliferation, migration, and apoptosis (Bae et al., 2001). Notably, *mir-135-a-1* has been reported to target *EGFR*, inhibiting cancer cell growth and migration (Xu et al., 2016). These findings suggest a possible interplay between *Caspase-7*, *mir-135-a-1*, and *EGFR*, underscoring the need for further investigation into their combined effect on cancer patient survival.

Table 6: Supporting evidence for synergistic interactions discovered by SynFS on the TCGA dataset.

View 1 (RPPA)	View 2 (miRNA)	Reference
Caspase-7	*mir-135a-1*	(Bae et al., 2001) (Xu et al., 2016)
Caspase-7	*mir-215*	(Su et al., 2015)
Myosin-iia	*mir-3199-1*	(Even-Ram et al., 2007) (Toro et al., 2018) (Cava et al., 2016)
Myosin-iia	*mir-320c-1*	(Liang et al., 2021)

6.5. Real-World Experiments: PBMC

Dataset Description. We analyze the distinction between two sub-populations of T-cells, specifically naive and regulatory T-cells, from purified populations of peripheral blood monocytes (PBMCs) using single-cell RNA (Zheng et al., 2017). The PBMC dataset comprises 105,868 samples, each characterized by 21,932 genes. For our analysis, we focus on 6,444 samples obtained from CD4 and CD8 T-cells, which play crucial roles in HIV-1 infection (Streeck et al., 2009). To construct a multi-view setting, we partition the RNA features into two distinct views based on their chromosome locations, i.e., chromosome 1 with $p_1 = 360$ genes and chromosome 2 $p_2 = 204$ genes. The choice of chromosomes is based on the presence of genes that demonstrate strong relevance to the CD4 and CD8 determinant LK2 protein (Hernández-Hoyos et al., 2000; Stelzer et al., 2016).

Quantitative Analysis. We select the top 22 features from each view based on the maximum difference between $SI(\mathcal{S})$ and $SI(\mathcal{N})$, while ensuring a high AUROC in the validation set. As shown in Table 5, SynFS outperforms the MLP trained with the entire features while utilizing merely 7% of the overall features. Furthermore, SynFS demonstrates competitive discriminative capability, slightly trailing behind CompFS. Notably, our method is the only model capable of accurately distinguishing synergistic and non-synergistic interactions. Furthermore, the result suggests the potential of SynFS to extend its applicability to single-view data when supplemented with expert curation of defining meaningful views.

Qualitative Analysis. We provide supporting evidence for the most frequently discovered synergistic features (more than 7 times out of 10 iterations) as summarized in Table 7. (Zhang et al., 2021) highlight the importance of epigenetic regulators like *RNF2* in regulatory T-cell development. We can hypothesize that *RNF2*'s epigenetic modifications could directly or indirectly influence genes involved in ribosome biogenesis, like *TTC31*, thus impacting the assembly of multiprotein complexes (Bakhshalizadeh et al., 2023) critical for T cell differentiation. This is supported by the broader research on the role of ribosome biogenesis in T-cell activation and differentiation, as seen in (Galloway et al., 2021).

Table 7: Supporting evidence for synergistic interactions discovered by SynFS on the PBMC dataset.

View 1 (Chromosome 1)	View 2 (Chromosome 2)	Reference
RNF2	*TTC31*	(Zhang et al., 2021) (Bakhshalizadeh et al., 2023) (Galloway et al., 2021)
HAX1	*PROC*	(Fan et al., 2022) (Cai et al., 2024) (Healy et al., 2018)

7. Conclusion

In this paper, we introduce SynFS, a novel multi-view feature selection method that is capable of discovering synergistic interactions. Our experiments demonstrate that SynFS can discover complex relationships between views that have been previously overlooked by traditional feature selection methods. Additionally, we validate the potential of our model for facilitating scientific discoveries in real-world applications by corroborating the discovered synergistic features with supporting medical and scientific literature.

Acknowledgements

We thank anonymous reviewers for many insightful comments and suggestions. CK and CL were supported through the IITP grant funded by the Korea government(MSIT) (No. 2021-0-01341, AI Graduate School Program, CAU).

Impact Statement

Our paper focuses on identifying features with synergistic interactions across different views in a multi-view setting. Our work goes beyond traditional feature selection methods by discovering features that exhibit greater predictive power when considered together with features from other views. In doing so, this paves the way for a more comprehensive understanding of the underlying interactions between features across multiple views, revealing how they collaborate to influence the outcome – something the conventional (single-view) feature selection methods cannot achieve.

For practical implication, this interaction discovery can be utilized in personalized treatment by targeting the factors that influence the disease. For example, immune checkpoint therapy (ICT) is a promising cancer treatment that helps the immune system recognize and attack tumors (Sammut et al., 2022). Analyzing a patient's tumor DNA (genomics) along with gene expression data (transcriptomic) could provide additional insights to identify patients who might not benefit from ICT despite having seemingly favorable mutations and provide alternative treatment.

References

Acharya, S., Cui, L., and Pan, Y. Multi-view feature selection for identifying gene markers: a diversified biological data driven approach. *BMC bioinformatics*, 21:1–31, 2020.

Agarap, A. F. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

Al-Khayyat, W., Pirkkanen, J., Dougherty, J., Laframboise, T., Dickinson, N., Khaper, N., Lees, S. J., Mendonca, M. S., Boreham, D. R., Tai, T. C., et al. Overexpression of fra1 (fosl1) leads to global transcriptional perturbations, reduced cellular adhesion and altered cell cycle progression. *Cells*, 12(19):2344, 2023.

Aronica, S. M., Kraus, W. L., and Katzenellenbogen, B. S. Estrogen action via the camp signaling pathway: stimulation of adenylate cyclase and camp-regulated gene transcription. *Proceedings of the National Academy of Sciences*, 91(18):8517–8521, 1994.

Bae, S. S., Choi, J. H., Oh, Y. S., Perry, D. K., Ryu, S. H., and Suh, P.-G. Proteolytic cleavage of epidermal growth factor receptor by caspases. *FEBS letters*, 491(1-2):16–20, 2001.

Bakhshalizadeh, S., Hock, D. H., Siddall, N. A., Kline, B. L., Sreenivasan, R., Bell, K. M., Casagranda, F., Kamalanathan, S., Sahoo, J., Narayanan, N., et al. Deficiency of the mitochondrial ribosomal subunit, mrpl50, causes autosomal recessive syndromic premature ovarian insufficiency. *Human Genetics*, 142(7):879–907, 2023.

Barabási, A.-L. Network medicine—from obesity to the "diseasome", 2007.

Benson, M. Clinical implications of omics and systems medicine: focus on predictive and individualized treatment. *Journal of internal medicine*, 279(3):229–240, 2016.

Binato, R., Corrêa, S., Panis, C., Ferreira, G., Petrone, I., da Costa, I. R., and Abdelhay, E. Nrip1 is activated by c-jun/c-fos and activates the expression of pgr, esr1 and ccnd1 in luminal a breast cancer. *Scientific Reports*, 11 (1):21159, 2021.

Bunnik, E. M. and Le Roch, K. G. An introduction to functional genomics and systems biology. *Advances in wound care*, 2(9):490–498, 2013.

Buttacavoli, M., Di Cara, G., D'Amico, C., Geraci, F., Pucci-Minafra, I., Feo, S., and Cancemi, P. Prognostic and functional significant of heat shock proteins (hsps) in breast cancer unveiled by multi-omics approaches. *Biology*, 10 (3):247, 2021.

Cai, W.-F., Jiang, L., Liang, J., Dutta, S., Huang, W., He, X., Wu, Z., Paul, C., Gao, X., Xu, M., et al. Hax1-overexpression augments cardioprotective efficacy of stem cell-based therapy through mediating hippo-yap signaling. *Stem Cell Reviews and Reports*, pp. 1–18, 2024.

Cava, C., Colaprico, A., Bertoli, G., Bontempi, G., Mauri, G., and Castiglioni, I. How interacting pathways are regulated by mirnas in breast cancer subtypes. *BMC bioinformatics*, 17:111–133, 2016.

Chen, J., Song, L., Wainwright, M., and Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*, pp. 883–892. PMLR, 2018.

Chen, M. and Denoyer, L. Multi-view generative adversarial networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 10*, pp. 175–188. Springer, 2017.

Dodge, K. L., Khouangsathiene, S., Kapiloff, M. S., Mouton, R., Hill, E. V., Houslay, M. D., Langeberg, L. K., and Scott, J. D. makap assembles a protein kinase a/pde4 phosphodiesterase camp signaling module. *The EMBO journal*, 20(8):1921–1930, 2001.

Even-Ram, S., Doyle, A. D., Conti, M. A., Matsumoto, K., Adelstein, R. S., and Yamada, K. M. Myosin iia regulates cell motility and actomyosin–microtubule crosstalk. *Nature cell biology*, 9(3):299–309, 2007.

Fan, Y., Murgia, M., Linder, M. I., Mizoguchi, Y., Wang, C., Łyszkiewicz, M., Zitara, N., Liu, Y., Frenz, S., Sciuccati, G., et al. Hax1-dependent control of mitochondrial proteostasis governs neutrophil granulocyte differentiation. *The Journal of Clinical Investigation*, 132(9), 2022.

Galloway, A., Kaskar, A., Ditsova, D., Atrih, A., Yoshikawa, H., Gomez-Moreira, C., Suska, O., Warminski, M., Grzela, R., Lamond, A. I., et al. Upregulation of rna cap methyltransferase rnmt drives ribosome biogenesis during t cell activation. *Nucleic Acids Research*, 49(12): 6722–6738, 2021.

Goyal, A., Morvant, E., Germain, P., and Amini, M.-R. Multiview boosting by controlling the diversity and the accuracy of view-specific voters. *Neurocomputing*, 358: 81–92, 2019.

Hagan, C. R., Daniel, A. R., Dressing, G. E., and Lange, C. A. Role of phosphorylation in progesterone receptor signaling and specificity. *Molecular and cellular endocrinology*, 357(1-2):43–49, 2012.

Healy, L. D., Rigg, R. A., Griffin, J. H., and McCarty, O. J. Regulation of immune cell signaling by activated protein c. *Journal of leukocyte biology*, 103(6):1197–1203, 2018.

Hernández-Hoyos, G., Sohn, S. J., Rothenberg, E. V., and Alberola-Ila, J. Lck activity controls cd4/cd8 t cell lineage commitment. *Immunity*, 12(3):313–322, 2000.

Ho, T. K. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pp. 278–282. IEEE, 1995.

Imrie, F., Norcliffe, A., Liò, P., and van der Schaar, M. Composite feature selection using deep ensembles. *Advances in Neural Information Processing Systems*, 35: 36142–36160, 2022.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Jaccard, P. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

Jackson, M., Marks, L., May, G. H., and Wilson, J. B. The genetic basis of disease. *Essays in biochemistry*, 62(5): 643–723, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kira, K. and Rendell, L. A. A practical approach to feature selection. In *Machine learning proceedings 1992*, pp. 249–256. Elsevier, 1992.

Kohavi, R. and John, G. H. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, C., Imrie, F., and van der Schaar, M. Self-supervision enhanced feature selection with correlated gates. In *International Conference on Learning Representations*, 2021.

Li, Y., Chen, C.-Y., and Wasserman, W. W. Deep feature selection: theory and application to identify enhancers and promoters. *Journal of Computational Biology*, 23(5): 322–336, 2016.

Li, Y., Tan, Y., Wen, L., Xing, Z., Wang, C., Zhang, L., Wu, K., Sun, H., Li, Y., Lei, Q., et al. Overexpression of birc6 driven by egf-jnk-hectd1 signaling is a potential therapeutic target for triple-negative breast cancer. *Molecular Therapy-Nucleic Acids*, 26:798–812, 2021.

Liang, Y., Li, S., and Tang, L. Microrna 320, an anti-oncogene target mirna for cancer therapy. *Biomedicines*, 9(6):591, 2021.

Lin, K. Z. and Zhang, N. R. Quantifying common and distinct information in single-cell multimodal data with tilted canonical correlation analysis. *Proceedings of the National Academy of Sciences*, 120(32):e2303647120, 2023.

Lindenbaum, O., Salhov, M., Averbuch, A., and Kluger, Y. L0-sparse canonical correlation analysis. In *International Conference on Learning Representations*, 2021.

Liu, H., Setiono, R., et al. A probabilistic approach to feature selection-a filter solution. In *ICML*, volume 96, pp. 319–327, 1996.

Liu, X., Sun, S. Q., Hassid, A., and Ostrom, R. S. camp inhibits transforming growth factor-β-stimulated collagen synthesis via inhibition of extracellular signal-regulated kinase 1/2 and smad signaling in cardiac fibroblasts. *Molecular pharmacology*, 70(6):1992–2003, 2006.

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

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

McGill, W. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.

Ogiwara, K., Hoyagi, M., and Takahashi, T. A central role for camp/epac/rap/pi3k/akt/creb signaling in lh-induced follicular pgr expression at medaka ovulation. *Biology of reproduction*, 105(2):413–426, 2021.

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications*, 7(1): 11479, 2016.

Petretto, E., Liu, E. T., and Aitman, T. J. A gene harvest revealing the archeology and complexity of human disease. *Nature Genetics*, 39(11):1299–1301, 2007.

Pfeifer, B., Baniecki, H., Saranti, A., Biecek, P., and Holzinger, A. Multi-omics disease module detection with an explainable greedy decision forest. *Scientific Reports*, 12(1):16857, 2022.

Sammut, S.-J., Crispin-Ortuzar, M., Chin, S.-F., Provenzano, E., Bardwell, H. A., Ma, W., Cope, W., Dariush, A., Dawson, S.-J., Abraham, J. E., et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601(7894):623–629, 2022.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, 54(1):1–30, 2016.

Streeck, H., Jolin, J. S., Qi, Y., Yassine-Diab, B., Johnson, R. C., Kwon, D. S., Addo, M. M., Brumme, C., Routy, J.-P., Little, S., et al. Human immunodeficiency virus type 1-specific cd8+ t-cell responses during primary infection are major determinants of the viral set point and loss of cd4+ t cells. *Journal of virology*, 83(15):7641–7648, 2009.

Su, Z., Yang, Z., Xu, Y., Chen, Y., and Yu, Q. Micrornas in apoptosis, autophagy and necroptosis. *Oncotarget*, 6(11): 8474, 2015.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Toro, L. E. N., Wang, Y., Condeelis, J. S., Jones, J. G., Backer, J. M., and Bresnick, A. R. Myosin-iia heavy chain phosphorylation on s1943 regulates tumor metastasis. *Experimental cell research*, 370(2):273–282, 2018.

Vidal, M., Cusick, M. E., and Barabási, A.-L. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

Wang, J., Xiu, J., Baca, Y., Battaglin, F., Arai, H., Kawanishi, N., Soni, S., Zhang, W., Millstein, J., Salhia, B., et al. Large-scale analysis of kmt2 mutations defines a distinctive molecular subset with treatment implication in gastric cancer. *Oncogene*, 40(30):4894–4905, 2021.

Wang, X., Gérard, C., Thériault, J.-F., Poirier, D., Doillon, C. J., and Lin, S.-X. Synergistic control of sex hormones by 17β-hsd type 7: a novel target for estrogen-dependent breast cancer. *Journal of molecular cell biology*, 7(6): 568–579, 2015.

Wu, J., Dent, P., Jelinek, T., Wolfman, A., Weber, M. J., and Sturgill, T. W. Inhibition of the egf-activated map kinase signaling pathway by adenosine 3, 5-monophosphate. *Science*, 262(5136):1065–1069, 1993.

Xu, B., Tao, T., Wang, Y., Fang, F., Huang, Y., Chen, S., Zhu, W., and Chen, M. hsa-mir-135a-1 inhibits prostate cancer cell growth and migration by targeting egfr. *Tumor Biology*, 37(10):14141–14151, 2016.

Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates. In *International Conference on Machine Learning*, pp. 10648–10659. PMLR, 2020.

Yang, J., Lindenbaum, O., and Kluger, Y. Locally sparse neural networks for tabular biomedical data. In *International Conference on Machine Learning*, pp. 25123–25153. PMLR, 2022.

Yang, J., Lindenbaum, O., Kluger, Y., and Jaffe, A. Multimodal differentiable unsupervised feature selection. In *Uncertainty in Artificial Intelligence*, pp. 2400–2410. PMLR, 2023.

Zhang, R., Nie, F., Li, X., and Wei, X. Feature selection with multi-view data: A survey. *Information Fusion*, 50: 158–167, 2019.

Zhang, W. and Liu, H. T. Mapk signal pathways in the regulation of cell proliferation in mammalian cells. *Cell research*, 12(1):9–18, 2002.

Zhang, Z., Luo, L., Xing, C., Chen, Y., Xu, P., Li, M., Zeng, L., Li, C., Ghosh, S., Della Manna, D., et al. Rnf2 ablation reprograms the tumor-immune microenvironment and stimulates durable nk and cd4+ t-cell-dependent antitumor immunity. *Nature cancer*, 2(10):1018–1038, 2021.

Zhao, L.-J., Hua, X., He, S.-F., Ren, H., and Qi, Z.-T. Interferon alpha regulates mapk and stat1 pathways in human hepatoma cells. *Virology journal*, 8(1):1–7, 2011.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

A. Appendix

A.1. Conditional Information and Interaction

Synergistic interaction between X, Z about Y is defined by comparing the MI between marginal, i.e., $I(Y;X)$ and conditional, i.e., $I(Y;X|Z)$ as explained in 3.2. This could be transformed into the comparison between the MI of the joint, i.e., $I(Y;X,Z)$, and the sum of marginals $I(Y;X)+I(Y;Z)$ applying chain rule as follows:

$$
\begin{aligned}
& I(Y;X|Z) > I(Y;X) \\
\Leftrightarrow\ & H(Y|Z) - H(Y|X,Z) > H(Y) - H(Y|X) \\
\Leftrightarrow\ & -H(Y|X,Z) > H(Y) - H(Y|X) - H(Y|Z) \\
\Leftrightarrow\ & H(Y) - H(Y|X,Z) > H(Y) - H(Y|X) + H(Y) - H(Y|Z) \\
\Leftrightarrow\ & I(Y;X,Z) > I(Y;X) + I(Y;Z)
\end{aligned}
$$

A.2. From the Problem of Maximizing Mutual Information to Minimizing Conditional Entropies

To identify synergistic features corresponding to the definition (4.2), we convert the maximizing(minimizing) MI terms into minimizing(maximizing) conditional entropy. We will show the derivation with 2 views for easier understanding, which can be easily extended to V views.

$$
\begin{aligned}
& \arg\max_S I(Y;\mathbf{X}^1_{\mathbf{g}_S},\mathbf{X}^2_{\mathbf{g}_S}) - I(Y,\mathbf{X}^1_{\mathbf{g}_S}) - I(Y,\mathbf{X}^2_{\mathbf{g}_S}) \\
\Leftrightarrow\ & \arg\max_S H(Y) - H(Y|\mathbf{X}^1_{\mathbf{g}_S},\mathbf{X}^2_{\mathbf{g}_S}) + H(Y|\mathbf{X}^1_{\mathbf{g}_S}) + H(Y|\mathbf{X}^2_{\mathbf{g}_S}) \\
\Leftrightarrow\ & \arg\min p(Y|\mathbf{X}^1_{\mathbf{g}_S},\mathbf{X}^2_{\mathbf{g}_S}) \log p(Y|\mathbf{X}^1_{\mathbf{g}_S},\mathbf{X}^2_{\mathbf{g}_S}) - p(Y|\mathbf{X}^1_{\mathbf{g}_S}) \log p(Y|\mathbf{X}^1_{\mathbf{g}_S}) - p(Y|\mathbf{X}^2_{\mathbf{g}_S}) \log p(Y|\mathbf{X}^2_{\mathbf{g}_S}) \\
\Leftrightarrow\ & \arg\min \mathbb{E} \log \frac{p(Y|\mathbf{X}^1_{\mathbf{g}_S},\mathbf{X}^2_{\mathbf{g}_S})}{p(Y|\mathbf{X}^1_{\mathbf{g}_S}) p(Y|\mathbf{X}^2_{\mathbf{g}_S})}
\end{aligned}
\tag{12}
$$

A.3. Pseudo-code of SynFS

Algorithm 1 Pseudo-code of SynFS

Input : $\mathcal{D} = \{\bar{\mathbf{x}}_i, \mathbf{y}_i\}_{i=1}^n$, selection regularization λ_S, λ_N,
mutual exclusivity hyperparameter α, mini-batch size n_{mb}, learning rate η,
Output : $\boldsymbol{\mu}_S, \boldsymbol{\mu}_N$
Initialize parameters $\{\boldsymbol{\mu}_S^v, \boldsymbol{\mu}_N^v\}_{v=1}^V, \phi_S, \phi_N, \phi_A$
while Converge **do**
  Sample a mini-batch from the dataset $(\bar{\mathbf{x}}_i, y_i)_{i=1}^{n_{mb}} \sim \mathcal{D}$
  for all $i = 1, \ldots, n_{mb}$ **do**
    Detach $\boldsymbol{\mu}_S, \boldsymbol{\mu}_N$ and calculate $\mathcal{L}_{pre}^{\phi_S,\phi_N}$

$$\mathcal{L}_{pre}^{\phi_S,\phi_N} \leftarrow \ell(y_i, f_{\phi_S}(\bar{\mathbf{x}}_{\mathbf{g}_S i})) + \sum_{v=1}^{V} \ell(y_i, f_{\phi_S}(\mathbf{x}_{\mathbf{g}_S i}^v)) + \ell(y_i, f_{\phi_N}(\bar{\mathbf{x}}_{\mathbf{g}_N i})) + \sum_{v=1}^{V} \ell(y_i, f_{\phi_N}(\mathbf{x}_{\mathbf{g}_N i}^v)) \tag{13}$$

    Update $\phi_S, \phi_N \leftarrow \phi_S - \eta\nabla_{\phi_S} \mathcal{L}_{pre}^{\phi_S,\phi_N}, \quad \phi_N - \eta\nabla_{\phi_N} \mathcal{L}_{pre}^{\phi_S,\phi_N}$

    Detach ϕ_N, ϕ_S and calculate $\mathcal{L}_{inf}^{\boldsymbol{\mu}_S,\boldsymbol{\mu}_N,\phi_A}$

$$\mathcal{L}_{inf}^{\boldsymbol{\mu}_S,\boldsymbol{\mu}_N,\phi_A} \leftarrow \ell(y_i, f_{\phi_A}(\bar{\mathbf{x}}_{Ai}) \tag{14}$$

    Update $\boldsymbol{\mu}_S, \boldsymbol{\mu}_N, \phi_A \leftarrow \boldsymbol{\mu}_S - \eta\nabla_{\boldsymbol{\mu}} \mathcal{L}_{inf}^{\boldsymbol{\mu}_S,\boldsymbol{\mu}_N,\phi_A}, \quad \boldsymbol{\mu}_N - \eta\nabla_{\boldsymbol{\mu}_S} \mathcal{L}_{inf}^{\boldsymbol{\mu}_S,\boldsymbol{\mu}_N,\phi_A}, \quad \phi_A - \eta\nabla_{\phi_A} \mathcal{L}_{inf}^{\boldsymbol{\mu}_S,\boldsymbol{\mu}_N,\phi_A}$

    Detach $\boldsymbol{\mu}_N, \phi_S, \phi_N, \phi_A$ and calculate $\mathcal{L}_{sel}^{\boldsymbol{\mu}_S}$

$$\mathcal{L}_{sel}^{\boldsymbol{\mu}_S} \leftarrow \ell(y_i, f_{\phi_S}(\bar{\mathbf{x}}_{\mathbf{g}_S i})) - \sum_{v=1}^{V} \ell(y_i, f_{\phi_S}(\mathbf{x}_{\mathbf{g}_S i}^v)) + \lambda_S \cdot \Phi(\boldsymbol{\mu}_S) \tag{15}$$

    Update $\boldsymbol{\mu}_S \leftarrow \boldsymbol{\mu}_S - \eta\nabla_{\boldsymbol{\mu}} \mathcal{L}_{sel}^{\boldsymbol{\mu}_S}$

    Detach $\boldsymbol{\mu}_S, \phi_S, \phi_N, \phi_A$ and calculate $\mathcal{L}_{sel}^{\boldsymbol{\mu}_N}$

$$\mathcal{L}_{sel}^{\boldsymbol{\mu}_N} \leftarrow -\ell(y_i, f_{\phi_N}(\bar{\mathbf{x}}_{\mathbf{g}_N i})) + \sum_{v=1}^{V} \ell(y_i, f_{\phi_N}(\mathbf{x}_{\mathbf{g}_N i}^v)) + \lambda_N \cdot \Phi(\boldsymbol{\mu}_N) + \alpha \cdot \text{sim}(\boldsymbol{g}_S, \boldsymbol{g}_N) \tag{16}$$

    Update $\boldsymbol{\mu}_N \leftarrow \nabla_{\boldsymbol{\mu}_N} \mathcal{L}_{sel}^{\boldsymbol{\mu}_N}$
  end for
end while

A.4. Extended related works

A few methods have addressed multi-view data without concatenation, capturing interactions between features from different views. However, these methods have limited ability to provide a more comprehensive understanding of how features interact across multiple views.

- **Greedy Decision Forest** (Pfeifer et al., 2022) selects multi-view features by leveraging graph knowledge to identify the smallest feature set that maximizes discriminative performance across multiple views. Initially, it samples features from diverse views to construct a tree based on random walks. Subsequently, the next tree is refined using a greedy algorithm. The process incorporates feature interaction through graph knowledge, constraining features within the same tree to be neighboring nodes in a graph.
- **UMVMO-select** (Acharya et al., 2020) constructs two views from diverse biological data, including gene ontology, protein interaction data, and protein sequences. The first view represents gene-gene dissimilarity, determined by pairwise correlation distances, while the second view captures protein similarity based on protein sequences. UMVMO-select solves the gene clustering problem by utilizing these two views and focuses on selecting the most frequently used genes.

Methods	Feature Selection	Multi-View	Interaction
CompFS (Imrie et al., 2022)	✓	✗	Regularization
Greedy (Pfeifer et al., 2022)	✓	✓	Graph Knowledge
UMVMO-select (Acharya et al., 2020)	✓	✓	Graph Knowledge
l_0**-DCCA** (Lindenbaum et al., 2021)	✓	✓	✗
MMDUFS (Yang et al., 2023)	✓	✓	✗
Tilted-CCA (Lin & Zhang, 2023)	✗	✓	✗

Table 8: Comparision table of related works.

A.5. Sensitivity Analysis

We have conducted the sensitivity analysis on the hyperparameters λ_S and λ_N on the Synthetic dataset (**Syn1**), which are the most important coefficients as they control the number of selected features with synergistic and non-synergistic interactions, respectively. Tables 9 and 10 present the feature selection performance for both synergistic (S-TPR, S-FDR), and non-synergistic (N-TPR, N-FDR) features, as well as for their union (J-index, TPR, FDR). We report the average score after 5 random iterations while varying the value of the target hyperparameter and fixing the other hyperparameters at their optimal values (marked with $*$). The results show that an overly strict constraint on the number of selected features will eventually lead our method to deprioritize important synergistic and non-synergistic features. Furthermore, the findings offer insights into determining the optimal value of λ_S on a smaller scale, which aligns with the intuition that synergistic interactions, being more intricate, are inherently harder to discover within the data structure.

λ_S	J-index ↑	TPR ↑	FDR ↓	S-TPR ↑	S-FDR ↓	N-TPR ↑	N-FDR ↓
0	0.62 ± 0.05	100.0 ± 0.0	62.4 ± 13.3	100.0 ± 0.0	76.0 ± 10.1	100.0 ± 0.0	0.0 ± 0.0
0.1	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0
0.25	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0
0.38*	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0
0.5	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0
0.6	0.80 ± 0.24	80.0 ± 24.4	0.0 ± 0.0	60.0 ± 48.9	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0
0.75	0.50 ± 0.00	50.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0

Table 9: The Feature Discovery Performance on the Synthetic dataset (Syn1) with varying λ_S

λ_N	J-index ↑	TPR ↑	FDR ↓	S-TPR ↑	S-FDR ↓	N-TPR ↑	N-FDR ↓
0	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.00	100.0 ± 0.0	0.0 ± 0.0
1.25*	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.00	100.0 ± 0.0	0.0 ± 0.0
2.07	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.00	100.0 ± 0.0	0.0 ± 0.0
4	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.00	100.0 ± 0.0	0.0 ± 0.0
8	1.00 ± 0.00	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.00	100.0 ± 0.0	0.0 ± 0.0
13	0.83 ± 0.20	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	13.3 ± 16.3	80.0 ± 24.4	0.0 ± 0.0
15	0.83 ± 0.20	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	13.3 ± 16.3	80.0 ± 24.4	0.0 ± 0.0
20	0.66 ± 0.17	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	26.6 ± 13.3	60.0 ± 20.0	0.0 ± 0.0

Table 10: The Feature Discovery Performance on the Synthetic dataset (Syn1) with varying λ_N

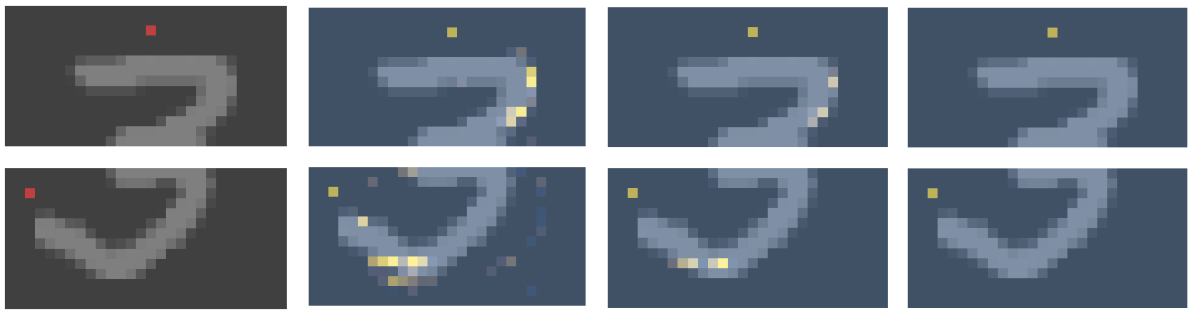
A.6. Semi-Synthetic MNIST

MNIST is a widely utilized dataset in feature selection research (Yamada et al., 2020; Imrie et al., 2022; Yang et al., 2022). Despite individual pixels lacking explicit significance, MNIST's standardized and centered format makes it an excellent dataset for visualizing feature selection outcomes. To investigate multi-view interactions, we split the MNIST images horizontally creating 2 views from the original 28×28 images to two 14×28 views. This allows for the analysis of the synergistic and non-synergistic features across views for digit classification. We compare the selected features of SynFS with CompFS and STG. For these single-view benchmark methods, we utilize the original 28×28 images and present the results using a horizontal split for improved readability.

Noisy MNIST We validate the discovered synergistic features by SynFS utilizing the noisy pixels – i.e., one pixel in each view with XOR and XNOR interactions to generate artificial labels. Figure 3 demonstrates that increasing regularization, λ_S, progressively isolates the synergistic features (i.e., the noisy pixels). Ultimately, our method pinpoints the two biases that are sufficient for perfect classification, showing the utility of SynFS in multi-view synergistic feature selection. To further validate the capability of SynFS to find the synergistic features in MNIST, we synthesize the MNIST data with synergistic features that the label is distinguishable with two features in each view. We designate a small bias in the range [0, 0.5) as TRUE and a large bias in the range [0.5, 1.0) as FALSE, and apply the exclusive generating rule as Table 11. For instance, if the initial random bias in View 1 is 0.35 and the image label is 0, the subsequent bias in View 2 is generated within the range [0.5, 1.0). Please see Table 11 for different scenarios for the data generation process.

Label	Relationship	View 1	View 2
0	XOR	True	False
		False	True
1	XNOR	True	True
		False	False

Table 11: Multi-view Noisy MNIST bias generating rule.



(a) Biases (b) $\lambda_S = 0.15$ (c) $\lambda_S = 0.5$ (d) $\lambda_S = 2.5$

Figure 3: Discovered synergistic features with varying λ_S for the MNIST dataset with injected biases (colored in red). The underlying digit '3' is used for better illustration.

A.7. METABRIC

To investigate the consistency of selected features across different random training/validation/testing splits, we analyzed the top 20 features from each view after 10 runs. Table 12 summarizes these results, including the total number of selected features and the number of features discovered more than 5 times.

View	Interaction	Total Number of Discovered Features	Frequently Discovered Features
Gene Expression	**Synergistic**	26	9
	Non-Synergistic	21	10
Mutation	**Synergistic**	45	7
	Non-Synergistic	21	9

Table 12: The number of selected features for the METABRIC dataset.

A.8. Experimental Details

Benchmark Implementation. For our experiments, we leveraged publicly available implementations of benchmark feature selection methods, ranging from traditional approaches like Random Forest (RForest) (Ho, 1995) to advanced methods such as STG (Yamada et al., 2020) and CompFS (Imrie et al., 2022). The hyper-parameters are chosen to maximize AUROC performance based on the validation set and detailed hyperparameters for synthetic experiments are presented in Table 13, with additional insights on benchmark methodologies and specific implementation details outlined below:

- Random Forest[5] utilizes an ensemble of decision trees, renowned for its robust performance on tabular data with diverse feature scales. A feature is considered important if it ranks within the top indexes based on its feature importance score across high-performing trees. Key hyperparameters and their respective search range include the number of trees (70-300), max tree depth (3-30), top trees for consideration (5-30), max feature indexes from a single tree (5-30), and a significance threshold (0.05-0.5).
- STG[6] employs a deep neural network for embedded feature selection, utilizing Gaussian-based continuous relaxation to handle binary selection variables. The regularization parameter λ (referred to as Reg1) is critical for controlling feature selection sparsity, with a search range from 1×10^{-5} to 2.0.
- CompFS[7] is an ensemble model for embedded feature selection, adopting Hard Concrete continuous relaxation (Maddison et al., 2016). CompFS is designed to distinguish interactive features by constituting groups of selected features from each selector. That is, having more than two features in one group represents there exists an interactive relationship among them. We further re-categorize each group to be synergistic if more than one features from one group are selected from different views, and non-synergistic otherwise. Hyperparameters β and β_R control sparsity and group overlap (Reg1 and Reg2), with the number of groups (i.e., estimators) set to 5 for our synthetic data experiments. The search range for Reg1 and Reg2 spans from 0.0 to 5.0.

SynFS Hyperparameters. SynFS incorporates synergistic and non-synergistic selectors to discover features based on their interaction types. Parameters λ_S and λ_N adjust the sparsity for each interaction type (Reg1 and Reg2), while α is a hyperparameter to ensure the mutual exclusivity of selected features from different interactions. For predictors that take selected features as input, we utilize dropout(Srivastava et al., 2014) with dropout probability 0.5 and batch normalization(Ioffe & Szegedy, 2015) to regularize the network.

Throughout the experiments, datasets encompassing multiple views are employed to evaluate the feature interaction discovery from different views. When applying single-view feature selection benchmark methods, we integrate these multiple views into a consolidated dataset through concatenation. For Deep learning based methods, we use a backbone network with 2-3 hidden layers with hidden width from $\{32, 64, 128\}$, number of epochs from 50 to 500, and learning rate from 1×10^{-3} to 1×10^{-5}. We employed the ADAM (Kingma & Ba, 2014) optimizer with ReLu (Agarap, 2018) activation.

Semi-Synthetic MNIST Implementations. For SynFS, we set the hidden dimensions to [400, 200, 300], with λ_S = 0.4, λ_N = 4.0, and α = 0.09. The learning rates are 1×10^{-3} for predictors and 2×10^{-3} for selectors. CompFS employs hidden dimensions of [256, 256], which is trained with 4 learners, setting both β and β_R to 0.18 with a learning rate of 1×10^{-3}. STG follows the same hidden dimension with CompFS and is trained with a learning rate of 2×10^{-3}. To compare the features selected by each method, we visualize the most significant 28 features: 14 for each view in SynFS , 7 features for each learner in CompFS, and the entirety of the view for STG.

[5]Python package `scikit-learn`

[6]https://github.com/runopti/stg

[7]https://github.com/a-norcliffe/Composite-Feature-Selection

Dataset	Methods	Reg1 (β, λ_N, Top Index)	Reg2 (β_R, λ_S, Top Tree)	α	Hidden Width & Max Depth	Epoch	lr	Estimators	Threshold
SYN1	SynFS	2.07	0.38	2.44	[128, 128]	95	0.001	-	0.7
	CompFS	0.58	2.41	-	[32, 32]	180	0.004	5	0.7
	STG	0.012	-	-	[64, 64]	445	0.003	-	0.7
	RForest	19	10	-	16	-	-	178	0.14
SYN2	SynFS	1.07	0.1	-	[32, 32]	90	0.001	-	0.7
	CompFS	1.88	0.025	-	[128, 128]	90	0.002	5	0.7
	STG	0.077	-	-	[128, 128]	371	0.005	-	0.7
	RForest	20	8	-	9	-	-	177	0.123
SYN3	SynFS	1.0	0.21	0.25	[64, 64]	90	0.001	-	0.7
	CompFS	0.257	1.54	-	[32, 32]	185	0.004	5	0.7
	STG	0.014	-	-	[64, 64]	2322	0.0038	-	0.7
	RForest	7	7	-	14	-	-	100	0.06
SYN4	SynFS	8.0	0.78	1.5	[32, 32, 32]	55	0.001	-	0.7
	CompFS	0.76	3.17	-	[32, 32]	125	0.0026	5	0.7
	STG	0.082	-	-	[128, 128]	327	0.0016	-	0.7
	RForest	24	10	-	24	-	-	100	0.215

Table 13: Hyperparameters for synthetic experiments.

Dataset	Methods	Reg1 (β, λ_N, Top Index)	Reg2 (β_R, λ_S)	α	Hidden Width & Max Depth	Epoch	lr	Estimators	Batch size	Top-K / Threshold
METABRIC	SynFS	1.95	0.45	1.06	[64, 64, 64]	175	0.00115	-	76	10
	CompFS	2.15	2.0	-	[32, 32]	121	0.005	4	76	5
	STG	0.185	-	-	[20, 20]	220	0.0014	-	87	20
	RForest	5	-	-	20	-	-	85	-	4
	MLP	-	-	-	[20, 20]	290	0.0007	-	55	-
TCGA	SynFS	0.54	0.22	0.59	[64, 64]	200	0.0029	-	128	0.55
	CompFS	0.23	3	-	[32, 32]	150	0.0029	2	32	77
	STG	1.95	-	-	[128, 128]	150	0.002	-	64	154
	RForest	64	-	-	19	-	-	232	-	2
	MLP	-	-	-	[128, 128]	274	0.0004	-	107	-
PBMC	SynFS	2.64	0.417	0.987	[64, 64]	50	0.002	-	64	11
	CompFS	3.47	3.89	-	[64, 64]	200	0.0039	2	64	11
	STG	0.206	-	-	[64, 64]	200	0.004	-	32	22
	RForest	26	-	-	9	-	-	106	-	2
	MLP	-	-	-	[128, 128]	100	0.0048	-	64	-

Table 14: Hyperparameters for real-world datasets experiments.

A.9. Predictive Performance of Synthetic and Semi-Syntheitc experiments

In this experiment, we evaluate the predictive performance of SynFS and baselines using each feature selection method as a pre-processing method for both synthetic and MNIST semi-synthetic datasets. For each method, we perform feature selection as described in Table. 3 and Figure. 2 and train a separate MLP using selected features as input. Table. 15 illustrates the superior performance of our method in selecting the most discriminative features across all synthetic tasks. Notably, the predictive performance of our method is comparable to that of feature selection methods focusing only on discriminative power, such as STG, which cannot provide explanations about the synergistic feature interactions. Similarly, Table. 16 demonstrates that our method outperforms baseline methods in classifying "3" and "8" by focusing on discriminative features that contribute synergistically to classification performance. This highlights that, by selecting synergistic features, our novel objectives achieve the dual benefit of fostering a deeper understanding of the data between the different views and demonstrably improving prediction performance.

Dataset	Methods	AUROC ↑	AUPRC ↑	Accuracy ↑	F1 Score ↑
Syn1	SynFS	0.999 ± 0.00	0.998 ± 0.00	0.963 ± 0.00	0.948 ± 0.00
	CompFS	0.999 ± 0.00	0.998 ± 0.00	0.964 ± 0.00	0.949 ± 0.00
	STG	0.999 ± 0.01	0.998 ± 0.01	0.963 ± 0.00	0.948 ± 0.01
	RForest	0.998 ± 0.00	0.997 ± 0.00	0.963 ± 0.00	0.948 ± 0.00
Syn2	SynFS	0.827 ± 0.00	0.828 ± 0.00	0.756 ± 0.00	0.761 ± 0.00
	CompFS	0.815 ± 0.03	0.814 ± 0.03	0.746 ± 0.02	0.751 ± 0.02
	STG	0.826 ± 0.01	0.826 ± 0.01	0.756 ± 0.00	0.762 ± 0.00
	RForest	0.769 ± 0.00	0.769 ± 0.00	0.706 ± 0.00	0.710 ± 0.00
Syn3	SynFS	0.817 ± 0.00	0.816 ± 0.01	0.745 ± 0.00	0.742 ± 0.01
	CompFS	0.804 ± 0.02	0.803 ± 0.02	0.733 ± 0.02	0.732 ± 0.02
	STG	0.818 ± 0.00	0.816 ± 0.01	0.746 ± 0.00	0.744 ± 0.00
	RForest	0.757 ± 0.04	0.745 ± 0.05	0.699 ± 0.00	0.702 ± 0.00
Syn4	SynFS	0.915 ± 0.00	0.920 ± 0.00	0.825 ± 0.00	0.825 ± 0.01
	CompFS	0.916 ± 0.00	0.921 ± 0.00	0.825 ± 0.00	0.825 ± 0.00
	STG	0.915 ± 0.00	0.920 ± 0.00	0.824 ± 0.00	0.825 ± 0.00
	RForest	0.872 ± 0.00	0.856 ± 0.00	0.814 ± 0.00	0.811 ± 0.00

Table 15: Predictive Performance on the Synthetic Dataset.

Dataset	Methods	AUROC ↑	AUPRC ↑	Accuracy ↑	F1 Score ↑
MNIST	SynFS	0.999 ± 0.00	0.999 ± 0.00	0.993 ± 0.00	0.993 ± 0.00
	CompFS	0.976 ± 0.02	0.975 ± 0.02	0.929 ± 0.03	0.928 ± 0.03
	STG	0.995 ± 0.00	0.994 ± 0.00	0.971 ± 0.00	0.971 ± 0.00

Table 16: Predictive Performance on the MNIST Dataset.

A.10. Interaction discovery with Known Ground Truth

The key distinction between the baselines and SynFS lies in their ability to *distinguish* the features with synergistic and non-synergistic interactions. An optimal model should proficiently discover informative features and categorize them into these two groups. To assess this, we employ the normalized Jaccard Index (J-Index), which evaluates whether the model successfully identifies and categorizes features into the correct groups. When ground truth feature interactions are known, we use the Jaccard Index which compares two sets with their union and intersection as below

$$\mathcal{J}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{17}$$

For methods that can distinguish interactions, we compare the ground truth feature sets, i.e., synergistic $\mathcal{S}$ and non-synergistic $\mathcal{N}$, with predicted synergistic and non-synergistic feature sets, $\hat{\mathcal{S}}, \hat{\mathcal{N}}$, respectively. For benchmark methods that select informative features without distinguishing the interactions, we compute *Group Similarity* (Imrie et al., 2022) which compares each ground truth feature set to the most similar set of features from the predicted sets. We will describe this more in detail with examples.

Jaccard Index: Metric for Methods with Feature Interactions. For SynFS, the Jaccard index is calculated as

$$\frac{1}{2}\big[\mathcal{J}(\mathcal{S}, \hat{\mathcal{S}}) + \mathcal{J}(\mathcal{N}, \hat{\mathcal{N}})\big]. \tag{18}$$

For CompFS (Imrie et al., 2022), we re-categorize the selected features when interactive features exist across views. For instance, when View 1 and View 2 have 10 dimensions each and if selected groups are $\{\{1, 13\}, \{2\}, \{4\}\}$, this means that $\{1, 13\}$ are synergistic features and non-synergistic features are $\{2, 4\}$.

Group Similarity: Metric for Methods with Feature Interactions. For other benchmark methods that are not specifically designed for differentiating interactions, we adopted the normalized Jaccard Index, called *Group Similarity* introduced in (Imrie et al., 2022). This metric compares the ground truth with the most similar set of features from the predicted groups and normalizes the score by the number G_{sim} of predictions as follows where 'T' and 'P' refer to the ground truth and the predicted features respectively:

$$G_{sim} = \frac{1}{\max(T, P)} \sum_{i=1}^{T} \max_{j \in [P]} \mathcal{J}(\mathcal{G}_i, \hat{\mathcal{G}}_j). \tag{19}$$

Please see Table 17 to help understand how the above metrics for interaction discovery are calculated depending on different methods. As demonstrated in the second and third examples in Table 17 even if a model successfully identifies informative features, misclassification into the wrong group can lead to a lower interaction discovery score.

Ground Truth		**Predicted**		**J-index**	**Predicted**	**Sum**	$\max(T, P)$	G_{sim}
Syn	**Non-Syn**	**Syn**	**Non-Syn**					
{1,3}	{2,4}	{1,3}	{2,4}	1	{2,4}, {1,3}	2	2	1
		{2,4}	{1,3}	0	{1,2}, {3,4}	2	2	1
		{1,2}	{3,4}	0.5	{1,2}, {3,4}, {5,6,7}	2	3	2/3

Table 17: A toy example of the similarity metrics.

A.11. Set Interaction Score

Set Interaction Scores (SI) measures the magnitude of synergistic interaction among selected features in a set by averaging the attribution of how each feature contributes to the interaction information. More specifically, given a set of features $\mathcal{S} = \mathcal{S}^1 \cup \cdots \cup \mathcal{S}^V$, we define the SI as the average contribution of each feature in the set, as follows:

$$SI(\mathcal{S}) = \frac{C}{|\mathcal{S}|} \sum_{v \in V} \sum_{d \in \mathcal{S}^v} \sum_{a \subset \mathcal{S}^v \setminus \{d\}} P(a + \{d\}) - P(a) \tag{20}$$

where C is a normalizing factor $\frac{|a|!(|S^v|-|a|-1)!}{|S^v|!}$. Here, $P(a)$ for $a \subset \mathcal{S}^v$ approximates an interaction score and is defined as $P(a) = \mathtt{F}(\mathcal{S}) - \sum_{i \in V \setminus v} \mathtt{F}(\mathcal{S}^i) - \mathtt{F}(a)$, where $\mathtt{F}$ symbolizes a performance metric, such as AUROC, which is computed only by taking input subset. To calculate the feature interaction attribution of a set, we compare the performance metrics with and without the inclusion of the specific feature. This combinatorial challenge necessitates considering all possible feature subsets within the interaction set.

For efficient computation, we randomly sample features from the identified interaction set. To ensure a robust evaluation of the performance across these variable feature subsets, we employ a neural network trained with a binary random mask, assigning a 70% probability for a mask value of zero. We further substantiate the effectiveness of the SI by assessing its performance on synthetic datasets with known ground truth interactions. Table 18 demonstrates the utility of SI by assigning higher scores for the known synergistic features ($SI(\mathcal{S})$) compared to the known non-synergistic features ($SI(\mathcal{N})$). This supports the effectiveness of SI in identifying features that work together for better prediction.

Dataset	$SI(\mathcal{S})$ $\uparrow$	$SI(\mathcal{N})$ $\downarrow$
Syn1	0.174 ± 0.00	-0.188 ± 0.00
Syn2	0.270 ± 0.00	-0.204 ± 0.00
Syn3	0.241 ± 0.00	-0.211 ± 0.00
Syn4	0.041 ± 0.00	N/A

Table 18: Set Interaction Score for Synthetic Dataset.