G^2M: A Generalized Gaussian Mirror Method to boost feature selection power

Hongyu Shen[1], **Zhizhen Zhao**[1]
[1] Department of Electrical and Computer Engineering
University of Illinois Urbana-Champaign
{hongyu2, zhizhenz}@illinois.edu

Abstract

Recent advances in false discovery rate (FDR)-controlled feature selection methods have improved reliability by effectively limiting false positives, making them well-suited for complex applications. A popular FDR-controlled framework called data splitting uses the "mirror statistics" to select features. However, we find that the unit variance assumption on mirror statistics could potentially limit the feature selection power. To address this, we generalize the mirror statistics in the Gaussian mirror framework and introduce a new approach called "generalized Gaussian mirror" (G^2M), which adaptively learns the variance and forms new test statistics. We demonstrate both theoretically and empirically that the proposed test statistics achieve higher power than those of Gaussian mirror and data splitting. Comparisons with other FDR-controlled frameworks on synthetic, semi-synthetic, and real datasets highlight the superior performance of the G^2M method in achieving higher power while maintaining FDR control. These findings suggest the potential for the G^2M method for practical applications in real-world problems. Code is available at: `https://github.com/skyve2012/G2M`.

1 Introduction

The complexity and high dimensionality of data in the era of big data make it challenging to develop a universal feature selection algorithm that performs well in diverse datasets [18]. An algorithm that works effectively for one type of data may fail for another. This variability complicates the verification of feature selection results, thereby hindering the practical use of selected features. Such challenges are especially prevalent in fields like healthcare, where there is no guarantee that identified genes or metabolites [8] correlate well with response variables (e.g., diseases), making downstream applications such as drug discovery even more difficult.

To address these challenges, recent research has focused on false discovery rate (FDR)-controlled feature selection methods [4, 8]. These methods limit the number of false positives during feature selection, providing a guarantee of the maximum number of incorrect selections. This property is particularly advantageous in ultra-high-dimensional settings, such as genetic or RNA-sequencing data.

Over the past decade, various methods have been proposed. Model-X knockoffs [4, 8] is a novel framework designed to select relevant features while controlling the FDR. It generates "knockoff" variables as references to the original design matrix and forms knockoff statistics. The symmetry-about-zero property of these statistics for null features enables FDR control under specified nominal levels. Although theoretical guarantees exist for Gaussian design matrices [8, 37], challenges remain for non-Gaussian settings due to the lack of theoretical assurances. Methods like Deep Knockoff [30], KnockoffGAN [14], and DeepDRK [34] have been proposed to enhance selection power while

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

controlling FDR in non-Gaussian data. However, as highlighted in Shen et al. [34], balancing reconstructability (which enhances power) and the swap property (which ensures FDR control) remains empirically difficult, and theoretical guarantees are still lacking.

Another line of work is the conditional randomization test (CRT), first introduced in Candès et al. [8]. CRT relies on conditional sampling for each feature given the others to compute p-value statistics, which are then used with the Benjamini-Hochberg (BH) procedure to control FDR. Unlike knockoff statistics, CRT avoids the reconstructability issue by sequentially considering individual features. However, its computational bottleneck arises from the need to repeatedly generate conditional samples, especially for non-Gaussian data where Markov Chain Monte Carlo (MCMC) techniques are required. This substantially increases the computational cost. Recent methods such as HRT [40], and Distillation-CRT [19] have sought to improve sampling accuracy and efficiency. Despite these advances, the inherent challenges of non-Gaussian data remain unresolved.

Distinct from these approaches, data splitting [9] and Gaussian mirror [45] introduce a "mirror statistics" with properties similar to knockoff statistics but without relying on the swap property. This avoids the trade-off identified in Shen et al. [34]. The key difference between the two methods lies in how the mirror statistics are constructed. Data splitting divides the design matrix and response into two parts, generating paired estimates for each feature, which are then used to compute the mirror statistics. In contrast, the Gaussian mirror introduces two Gaussian noise perturbations to create additional columns in the design matrix, replacing the original feature column [1]. These perturbed columns are used to compute the mirror statistics. However, on the one hand, the data-splitting approach is constrained by its unit-variance assumption for the mirror statistics. On the other hand, the Gaussian mirror paper does not provide a discussion on the statistical power of the proposed method.

To address these limitations, we generalize the variance assumptions in the mirror statistics introduced in the data-splitting framework and propose a generalized Gaussian mirror test statistics that considers the variance information of the fitting coefficients—an intermediate component that forms the mirror statistics. We demonstrate that this statistics is the entry-wise uniformly most powerful test statistics, leading to improved feature selection power compared to both the Gaussian mirror and data-splitting test statistics. We call the proposed method a "generalized Gaussian mirror" (G^2M) [2]. With experiments considering synthetic, semi-synthetic, and real datasets, we demonstrate the practical value of the proposed approaches over a set of benchmarking methods.

2 Related Work

2.1 Data Splitting

Data splitting is a feature selection method designed to control the FDR [9]. This method partitions the design matrix X and the response vector y into two disjoint datasets, denoted as (X^+, y^+) and (X^-, y^-). Separate models are then fitted to each dataset using the same approach (e.g., both use ordinary least squares), yielding the coefficients β_j^+ and β_j^- from two respective models for each feature x_j. For each pair (β_j^+, β_j^-), the method constructs a test statistics w_j, which is a function of these coefficients and reflects the importance of the corresponding feature x_j.

Notably, the statistics w_j is designed as a "mirror statistics," ensuring symmetry about 0 for null features ($j \in S_0$) while maintaining positive values for non-null features ($j \in S_1$). After calculating w_j for all features, the selection procedure identifies the set of significant features $\hat{S}_1 = \{j : w_j \geq \tau_q\}$, where the threshold τ_q is determined based on the desired FDR control.

$$\tau_q = \min_{t>0} \Big\{ t : \frac{1 + |\{j : w_j \leq -t\}|}{\max(1, |\{j : w_j \geq t\}|)} \leq q \Big\}. \tag{1}$$

This setting controls feature selection FDR level at q according to Theorem 2.1.

[1] It is noteworthy that the data-splitting (Section 2.1) and Gaussian mirror (Section 2.2) methods are closely related to the proposed G^2M method. We describe their methods individually and discuss the linkage to our proposed approach in Section 2.

[2] To the best of our knowledge, there are limited works concerning improvements of data splitting or Gaussian mirror. Ge et al. [12] considers extending an uneven data splitting with cox model. Dai et al. [10], Xing et al. [44] consider nonlinearity in test statistics. These are orthogonal to our focus, therefore omitted.

Theorem 2.1. *For a given set of mirror statistics $\{w_j\}_{j=1}^p$, the following properties hold:*

1. *If j is a null feature index, then for any real number t, $\mathbb{P}(w_j \geq t) = \mathbb{P}(w_j \leq -t)$.*
2. *If j is a non-null feature index, then $\mathbb{P}(w_j \geq 0) > \mathbb{P}(w_j \leq 0)$.*

Then, for a given nominal FDR level q, the feature selection set

$$\widehat{S} = \{j \in \{1, \ldots, p\} : w_j \geq \tau_q\}$$

controls the FDR, where τ_q is chosen via Eq. (1)*.*

Proof. Given the above two properties for the mirror statistics, we have $\mathbb{P}(w_j \geq t) = \mathbb{P}(w_j \leq -t)$ due to the symmetry-about-zero property of the mirror statistics w_j under the null distribution. As a result, the selection rule $\widehat{S} = \{j : w_j \geq \tau_q\}$ (equivalently, the denominator of Eq. (1)) selects features based on the level-q quantile on the right side of the null distribution, ensuring the FDR level is controlled at most q. □

2.2 Gaussian Mirror

The Gaussian mirror [45] is a different approach from the data splitting method given how β_j^+ and β_j^- are generated. Specifically, given an index j, the Gaussian mirror proposes two perturbed variables in place of the original x_j, resulting in a new design matrix $X^j = (x_j^+, x_j^-, X_{-j})$, where $x_j^+ = x_j + c_j z_j$, $x_j^- = x_j - c_j z_j$, $z_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and c_j is a scalar in $\mathbb{R}$. β_j^+ and β_j^- are the coefficient estimations for x_j^+ and x_j^-, respectively.

The Gaussian mirror adopts a similar notion of mirror statistics as the data splitting method to select features with controlled FDR. With $c_j = \frac{\|P_{\perp -j} x_j\|}{\|P_{\perp -j} z_j\|}$, Xing et al. [45] shows that β_j^+ and β_j^- are independent, allowing the choice of a specific form of mirror statistics $w_j = |\beta_j^+ + \beta_j^-| - |\beta_j^+ - \beta_j^-|$ to perform FDR-controlled feature selection at level q, according to Eq. (1). Here, $P_{\perp -j}$ denotes the projection in the null space of X_{-j}.

2.3 Connection between Data Splitting and Gaussian Mirror in the G^{2}M setting

In this paper, the proposed G^{2}M test statistics is based on a variance-generalized uniformly most powerful (UMP) statistics under the mirror statistics setting in the data splitting paper [9]. Specifically, it adopts the β_j^+ and β_j^- from the Gaussian mirror paper [45] to form a generalized mirror statistics (see Lemma. 3.5) over the one from the data splitting paper. The proposed statistics considers the variance information of β_j^+ and β_j^-, resulting in higher feature selection power compared to that of the data splitting and Gaussian mirror. More importantly, it is theoretically verifiable. In addition, We provide evidence on the no-constant variance of β_j^+ and β_j^- in Appendix A—the rationale on why such variance generalization is necessary.

3 Method

This section details the proposed approach in this paper. The arrangement is as follows:

- Section 3.2 considers a more general setup of Gaussian mirror variable x_j^+ and x_j^-. And provide the related form of β_j^+ and β_j^- as a weighted sum of the true β_j coefficient and the noise variables. We also provide two useful properties about β_j^+ and β_j^- whose proofs are omitted in the original Gaussian mirror paper [45].
- Section 3.3 presents the first major result (i.e., Lemma 3.5) on the UMP test statistics given the feature index j. The setup we consider is more realistic compared to the one considered in Dai et al. [9], resulting in higher test power.
- Section 3.4 introduces the second major result that proves the proposed G^{2}M test statistics (see Proposition 3.1) is a more powerful test statistics compared to the mirror statistics in Xing et al. [45], Dai et al. [9].

- In Section 3.5, we combine the results discussed in the previous sections and propose two algorithms—one exact algorithm that characterizes an ideal scenario, and one estimation algorithm that can be used in practice. Experiments in Section 4 are conducted w.r.t. the estimation version.

3.1 Notation

We first introduce the notation for the paper. Let $X \in \mathbb{R}^{n\times p}$ be the design matrix, $x_j = (x_{1j}, \ldots, x_{nj})^\top$ be the j-th column of X, and X_{-j} be the submatrix of X with the j-th column removed. n is the sample size and p is the feature dimension. Following the setting in Xing et al. [45], we assume that X is normalized such that $\sum_{i=1}^n x_{ij} = 0$ and $\|x_j\|_2 = n$, $j \in [1, \ldots, p]$. Let $y = (y_1, \ldots, y_n)^\top$ be the vector of n independent responses. We assume that the response variable y only depends on a subset of features $X_{S_1} = \{X_j : j \in S_1\}$, and the task of feature selection is to identify the set S_1. And the response y follows a linear model: $y = X\beta + \epsilon$, where $\beta = (\beta_1, \ldots, \beta_p)$ indicates the coefficients for the features x_j's. Let $S_0 = \{1, \ldots, p\} \setminus S_1$ be the index set of the null features. Let $p_0 = |S_0|$ and $p_1 = |S_1|$ be the number of the null and the relevant features, respectively.

3.2 A General form of Gaussian Mirror Coefficient Estimation

We propose to represent β_j^+ and β_j^- as a function of the true β_j and the noise vector ϵ, given a generalized Gaussian mirror setup that considers additional variables (i.e., d_j and q_j). This improves flexibility on the representation of β_j^+ and β_j^-. Details are included in Proposition 3.1, which helps us to understand the behavior of β_j^+ and β_j^- in terms of q_j, z_j, c_j, d_j, β_j and ϵ.

Proposition 3.1. *Consider a generalized Gaussian mirror setup where $x_j^+ = x_j + c_j \cdot z_j$ and $x_j^- = x_j - d_j \cdot q_j$, and z_j and q_j are n dimensional i.i.d. standard Gaussian. c_j and d_j are in $\mathbb{R}$. Then the least square regression coefficients for x_j^+ and x_j^- have the following forms:*

$$\beta_j^+ = \alpha \cdot \beta_j + \gamma^\top \cdot \epsilon, \quad \beta_j^- = \zeta \cdot \beta_j + \eta^\top \cdot \epsilon, \tag{2}$$

where ϵ stands for n-dimensional i.i.d. standard Gaussian vector. And α, γ, ζ, and η are functions of c_j, d_j, z_j, and q_j (see Appendix B.1 for full forms).

In addition to Proposition 3.1, we introduce three corollaries that reveal certain properties about β_j^+ and β_j^- and terms associated with them (see Eq. (2)). These properties are crucial in developing the estimation algorithm in Section 3.5.

Corollary 3.2. *Given $z_j = q_j$, $\alpha + \zeta = 1$, where α and ζ are defined in Proposition 3.1.*

Corollary 3.3. *Given $c_j = d_j$, and $z_j = q_j$, $\alpha = \zeta = 0.5$, where α and ζ are defined in Proposition 3.1.*

Specifically, Corollary 3.2 describes a property on the behavior of α and η that sum to one as long as the Gaussian mirror perturbation vectors $z_j = q_j$. Notice that this does not require $c_j = d_j$, indicating a more generalized setup beyond the Gaussian mirror. On the other hand Corollary 3.3 indicates that under the vanilla Gaussian mirror setup (see Section 2.2 and Xing et al. [45]), the mean of β_j^+ and β_j^- is 0.5β. Note that the result is only mentioned in Xing et al. [45] without any proof. Here we provide formal proof of this fact, complementing the original Gaussian mirror work. The proof of Corollary 3.2 and 3.3 can be found in Appendix B.

Corollary 3.4. *Var$(\beta_j^+) = \|\gamma\|_2^2$ and Var$(\beta_j^-) = \|\eta\|_2^2$.*

The proof is straightforward by realizing that ϵ is the only term that introduces randomness to β_j^+ and β_j^-. Hence Var(β_j^+) = Var$(\gamma^\top \cdot \epsilon) = \|\gamma\|_2^2$. Similar reasoning can be applied to Var(β_j^-).

3.3 Entry-wise Uniformly Most Powerful Mirror Statistics

According to Section 2.3 and Appendix A, it is clear that the variances of β_j^+ and β_j^- need not be 1. This violates the condition for the UMP mirror test statistics proposed in Dai et al. [9]. To consider a

more realistic setup, we present the first result (i.e., Lemma 3.5) that reveals a mirror test statistics that is a function of the variance of β_j^+ and β_j^-. We show that the proposed test statistics is UMP for every pair of β_j^+ and β_j^-. Note that we do not assume a universal distribution for β_j^+ and β_j^- across all j's. This flexibility directly leads to Theorem 3.6, which proves that the test statistics in Lemma 3.5 achieves higher power during feature selection compared to the existing mirror statistics [45, 9].

Lemma 3.5. *Suppose that the set of coefficients (β_j^+, β_j^-) is independent for all $j \in \{1, \ldots, p\}$. Furthermore, suppose:*

(a) For $j \in S_1$, the two coefficients β_j^+ and β_j^- follow $N(\omega, \sigma_a^2)$, $N(\omega, \sigma_b^2)$ independently. And $\omega \sim \delta \cdot Rademacher(0.5)$ ($\delta > 0$), where $\delta \in \mathbb{R}$ is the absolute magnitude of ω [3];

(b) For $j \in S_0$, the two coefficients β_j^+ and β_j^- follow $N(0, \sigma_a^2)$ and $N(0, \sigma_b^2)$ independently;

(c) $\frac{p_1}{p_0} \to r$ as $p \to \infty$.

Then the optimal choice of f in the mirror statistics (i.e., $\text{sign}(\beta_j^+ \beta_j^-) f(|\beta_j^+|, |\beta_j^-|)$) that yields the highest power follows the form:

$$f(a, b) = U \left[P_+ \exp(-S_-) + P_- \exp(-S_+) \right], \tag{3}$$

where $S_- = \frac{\delta(\delta - 2a)}{2\sigma_a^2} + \frac{\delta(\delta - 2b)}{2\sigma_b^2}$, $S_+ = \frac{\delta(\delta + 2a)}{2\sigma_a^2} + \frac{\delta(\delta + 2b)}{2\sigma_b^2}$, and

$$P_+ = \Phi\left(\frac{\delta}{\sigma_a}\right) \Phi\left(\frac{\delta}{\sigma_b}\right), \; P_- = \left[1 - \Phi\left(\frac{\delta}{\sigma_a}\right)\right] \left[1 - \Phi\left(\frac{\delta}{\sigma_b}\right)\right], \; U = \frac{1}{P_{E|H_1}} = \frac{1}{P_- + P_+},$$

where Φ is the cumulative density function for the standard normal distribution.

Lemma 3.5 reveals a closed-form test statistics that is UMP. This is used in developing the proposed feature selection algorithms in Section 3.5.

3.4 More Powerful Test Setting

Combining the results from Proposition 3.1 and Lemma 3.5, we present the following Theorem 3.6 that indicates a more powerful test statistics of the proposed over the existing ones in Xing et al. [45], Dai et al. [9].

Theorem 3.6. *The proposed test statistics $f(a, b)$ in Eq.* (3) *achieves the highest power compared to the Gaussian mirror test statistics: $|a + b| - |a - b|$ [45], and the data splitting test statistics: $\text{sign}(ab)(|a| + |b|)$ [9], under the same nominal FDR level q.*

In Section 4, we will empirically observe evidence that supports Theorem 3.6. And the proof of Theorem 3.6 can be found in Appendix B.

3.5 Algorithm

With all the ingredients introduced in the previous sections, we present two algorithms: 1. the exact algorithm (Algorithm 1) when one has access to the scale of the true coefficients for the nonnull β_j's: δ or δ_j if it is index-dependent; 2. the estimation algorithm (Algorithm 2) that essentially proposes an estimation of δ or δ_j, resulting a generalized likelihood ratio test statistics given the estimation. Note that the proposed algorithm considers the vanilla Gaussian mirror setup to take advantage of the nice properties indicated in Corollary 3.2 and 3.3. This means $c_j = d_j$ and $z_j = q_j$ for all j's. And $c_j = \frac{\|P_{\perp - j} x_j\|}{\|P_{\perp - j} z_j\|}$ following the setup in Xing et al. [45].

Algorithm 1 introduces the exact algorithm when true δ presents. This is rarely the case as often we do not have prior knowledge about true β_j's. To overcome this issue, we propose an algorithm to estimate δ, which is Algorithm 2.

Essentially, the estimation algorithm employs a k-means-based algorithm to find k potential modes for δ_j—a quantity that is assumed to be given in Algorithm 1, yet unavailable in practice. We

[3] We adopt a similar setting in Proposition 1 in Dai et al. [9] to define ω.

Algorithm 1 Exact algorithm

1: **Input:** Design matrix X, response y, nominal FDR level q and true scale δ or δ_j for the β_j, $j \in S_1$.
2: **Output:** $\hat{S}_1 = \{j \mid w_j \geq \tau_q\}$.
3: **for** $j = 1$ **to** p **do**
4: Sample $z_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$
5: Calculate $c_j = \frac{\|P_{\perp -j} x_j\|}{\|P_{\perp -j} z_j\|}$ and form $x_j^+ = x_j + c_j z_j$, $x_j^- = x_j - c_j z_j$.
6: Obtain the ordinary least squares estimator of $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$:

$$(\hat{\beta}_j^+, \hat{\beta}_j^-, \hat{\beta}_{-j}) = \arg \min_{\beta_j^+, \beta_j^-, \beta_{-j}} \|y - X_{-j}\beta_{-j} - x_j^+ \beta_j^+ - x_j^- \beta_j^-\|_2^2.$$

7: Calculate $\text{Var}(\beta_j^+) = \|\gamma\|_2^2$ and $\text{Var}(\beta_j^-) = \|\eta\|_2^2$.
8: Compute $w_j = \text{sign}(\beta_j^+ \beta_j^-) f(\beta_j^+, \beta_j^-)$, where $f(a, b)$ is specified in Eq. (3).
9: **end for**
10: Calculate the threshold given the FDR level q:

$$\tau_q = \min_{t>0} \Big\{ t : \frac{1 + |\{j : w_j \leq -t\}|}{\max(1, |\{j : w_j \geq t\}|)} \leq q \Big\}.$$

choose k-means given it has been widely considered as an efficient unsupervised algorithm due to no accessible to true values of δ_j. First, combining the results of Corollary 3.2 and 3.3, the algorithm obtains an unbiased estimate of the absolute value of β_j ($j \in (1, \ldots, p)$) in Step 7 of Algorithm 2. This is because α and ζ are both equal to 0.5 based on Corollary 3.2 and 3.3. The p calculated values represent possible values of δ. Then we utilize the k-means algorithm with k chosen via the silhouette score—a common technique for choosing the number of clusters. Note that the k identified modes include the case for the null β_j's, where $\delta = 0$, due to not knowing which variable is null and which is nonnull. Luckily, we can simply exclude the smallest mode (e.g., the closest one to zero), which corresponds to the mode for the null β_j's, resulting in $k - 1$ candidates. For every calculated pair of β_j^+ and β_j^-, $\hat{\delta}_j$ is chosen to be the closest mode among the $k - 1$ candidates, compared to the value $\frac{\beta_j^+ + \beta_j^-}{2}$. With Algorithm 2, we can perform feature selection without accessing the true δ information. In Section 4, all results about the proposed G^{2}M method are based on the estimation algorithm (Algorithm 2).

Algorithm 2 Estimation algorithm

1: **Input:** Design matrix X, response y, nominal FDR level q and the number of modes for δ: $k \leq p$.
2: **Output:** Algorithm 1(X, y, $\{\hat{\delta}_j\}_{j=1}^p$, q)
3: **for** $j = 1$ **to** p **do**
4: Sample $z_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$
5: Calculate $c_j = \frac{\|P_{\perp -j} x_j\|}{\|P_{\perp -j} z_j\|}$ and form $x_j^+ = x_j + c_j z_j$, $x_j^- = x_j - c_j z_j$.
6: Obtain the ordinary least squares estimator of $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$:

$$(\hat{\beta}_j^+, \hat{\beta}_j^-, \hat{\beta}_{-j}) = \arg \min_{\beta_j^+, \beta_j^-, \beta_{-j}} \|y - X_{-j}\beta_{-j} - x_j^+ \beta_j^+ - x_j^- \beta_j^-\|_2^2.$$

7: Calculate the value $v_j = \frac{|\beta_j^+ + \beta_j^-|}{2}$.
8: **end for**
9: Run k-means algorithm to find k modes given $\{v_j\}_{j=1}^p$, represented as $\{m_l\}_{l=1}^k$.
10: **for** $j = 1$ **to** p **do**
11: $\hat{\delta}_j = \min_{m_l} |v_j - m_l|, l \in (1, \ldots, k)$.
12: **end for**

4 Experiment

In this section, we evaluate the proposed G^{2}M approach using three types of datasets, following the experimental setup described in Shen et al. [34] given its comprehensibility. Specifically, we consider: 1. Fully synthetic datasets, where the design matrix X and the response y are generated from known distributions; 2. Semi-synthetic datasets, where the design matrix is extracted from real-world data, and the response is generated using a predefined model; 3. A real-world case study, where both X and y are fully unknown. The experimental details and corresponding results are summarized in Section 4.1, Section 4.2, and Section 4.3, respectively. In each section, we provide details on the benchmarking methods, the configuration of the dataset, and the results. All experiments were carried out on an IBM AC922 server with 2x 20 core IBM POWER9 CPU @ 2.4GHz.

Table 1: FDR and power with Gaussian design matrix X for $\rho = 0.6$ (left) and $\rho = 0.7$ (right). **Bold** entries indicate the highest power given controlled FDR at level 0.1. **Blue** for the second best, and Red for FDR> 0.1.

Method	$\rho = 0.6$						$\rho = 0.7$					
	OLS		Ridge		LASSO		OLS		Ridge		LASSO	
	FDR	Power	FDR	Power	FDR	Power	FDR	Power	FDR	Power	FDR	Power
CRT [8]	0.00	0.13	0.07	0.69	0.07	0.77	0.00	0.02	0.13	0.51	0.11	0.56
Distilled-CRT [19]	0.53	0.98	0.27	0.98	0.13	0.97	0.58	0.95	0.25	0.93	0.11	0.87
Gaussian Mirror [45]	0.09	0.54	0.05	0.67	0.04	0.74	0.05	0.18	0.04	0.51	0.23	0.48
Gaussian Mirror† [45]	0.10	0.61	0.05	0.71	0.04	0.76	0.13	0.49	0.06	0.64	0.04	0.61
Data Splitting [9]	0.00	0.00	0.06	0.66	0.04	0.50	0.00	0.00	0.06	0.56	0.05	0.41
Data Splitting† [9]	0.72	0.07	0.08	0.71	0.17	0.70	0.73	0.07	**0.07**	**0.66**	0.25	0.70
HRT [40]	0.01	0.23	0.00	0.21	0.00	0.22	0.01	0.10	0.00	0.10	0.00	0.10
Powerful Knockoff [37]	0.00	0.00	0.05	0.48	0.06	0.69	0.01	0.00	0.04	0.29	0.05	0.55
Powerful Knockoff† [37]	0.75	0.08	0.08	0.61	0.06	0.73	0.80	0.07	0.07	0.50	0.05	0.63
G^{2}M (ours)	**0.08**	**0.65**	**0.05**	**0.82**	**0.05**	**0.82**	**0.08**	**0.36**	**0.03**	**0.67**	**0.03**	**0.68**
G^{2}M† (ours)	**0.08**	**0.72**	**0.04**	**0.80**	**0.06**	**0.84**	**0.10**	**0.55**	**0.03**	**0.67**	**0.04**	**0.70**

4.1 Synthetic Data

To thoroughly evaluate performance, we adopt the experimental framework outlined in Shen et al. [34] to generate diverse synthetic datasets defined by (X, y). In this setup, $X \in \mathbb{R}^p$ represents dependent variables drawn from predefined distributions, and $y \in \mathbb{R}$ serves as the response variable. We model y using a linear response framework: $y \sim \mathcal{N}(X^T\beta, 1)$. The true underlying β is a p-dimensional vector with entries independently sampled from the distribution $\frac{p}{15\cdot\sqrt{n}} \cdot \text{Rademacher}(0.5)$. This choice differs from the commonly used scaling factor $\frac{p}{\sqrt{n}} \cdot \text{Rademacher}(0.5)$ [38, 22], enabling a more challenging evaluation configuration.

In addition to the Copula and Gaussian mixture models considered in Shen et al. [34] that quantifies the nonlinear design matrices, we include the Gaussian setting, exploring various correlation levels among the features. Since methods such as knockoffs [8] and the conditional randomization test (CRT) [8] have closed-form solutions in Gaussian settings, incorporating this setup offers valuable insights into the consistency between empirical results and theoretical guarantees. Due to space limit, we defer the introduction of these settings in Appendix D.

Benchmarking Methods & Settings [4]**:** Given the linear and nonlinear nature of the design matrix, we utilize two sets of benchmarking methods. The first set comprises methods that are theoretically proven to control the false discovery rate (FDR), including knockoff [8], conditional randomization test (CRT) [8], data splitting [9], Gaussian mirror [45], distilled-CRT [19], HRT [40], and powerful knockoff [37]. In addition, we also consider another rule to choose τ_q (Eq. (4)) according to Ren and Barber [29], to improve the power for certain applicable methods and mark them with "†":

$$\tau_q = \min_{t>0}\Big\{t : \frac{1+|\{j : w_j \le -t\}|}{\max(1, |\{j : w_j \ge t\}|)} \le q \quad \text{or} \sum_{j\in[p]} \mathbb{1}\{j : w_j \ge t\} < \frac{1}{q}\Big\}, \tag{4}$$

where "$\mathbb{1}$" is the indicator function. Essentially, the two equations differ only when $\min_{t>0}\{t : \frac{1+|\{j:w_j\le -t\}|}{\max(1,|\{j:w_j\ge t\}|)} \le q\} > \min_{t>0}\{t : \sum_{j\in[p]} \mathbb{1}\{j : w_j \ge t\} < \frac{1}{q}\}$. This means when q is small or

[4] We include discussion about computation complexity for non-deep-learning-based methods. We defer the discussion in Appendix C given this is not the major focus on this work

the fraction of non-nulls is small, then the first term usually finds a larger τ_q, in favor of the FDR control. In this case, the power is pretty low, as when τ_q is large, there are only limited non-nulls selected. Instead of choosing τ_q with the first term (e.g., Eq. (1)), the second term chooses a smaller τ_q, which results in higher power.

The second set focuses on deep learning-based methods designed to handle nonlinear design matrices, such as Deep Knockoff [30] [5], DDLK [38] [6], KnockoffGAN [14] [7], sRMMD [22] [8], and DeepDRK [34] [14].

Table 2: FDR and power with Copula and Gaussian Mixture design matrix X. **Bold** entries indicate the highest power with controlled FDR at level 0.1. **Blue** for the second best, and Red for FDR>0.1.

Method	Citation	Gaussian Mixture		Clayton & Exp.		Clayton & Gamma		Joe & Exp.		Joe & Gamma	
		FDR	Power	FDR	Power	FDR	Power	FDR	Power	FDR	Power
CRT	[8]	0.00	0.21	0.08	0.69	0.05	0.91	0.08	0.45	0.07	0.87
Distilled-CRT	[19]	0.17	0.22	0.06	0.43	0.06	0.36	0.05	0.26	0.04	0.26
Gaussian Mirror	[45]	0.05	0.83	0.07	0.54	0.07	0.89	0.06	0.34	0.08	0.86
Gaussian Mirror†	[45]	0.04	0.83	**0.07**	**0.70**	0.08	0.92	**0.09**	**0.52**	0.09	0.84
Data Splitting	[9]	0.07	0.72	0.07	0.52	0.07	0.79	0.06	0.28	0.08	0.76
Data Splitting†	[9]	0.09	0.76	0.09	0.62	0.08	0.81	0.10	0.49	0.09	0.78
HRT	[40]	0.00	0.18	0.01	0.15	0.01	0.50	0.01	0.09	0.02	0.50
Powerful Knockoff	[37]	0.08	0.62	0.03	0.17	0.05	0.38	0.04	0.11	0.07	0.39
Powerful Knockoff†	[37]	0.09	0.64	0.12	0.40	0.06	0.51	0.14	0.39	0.07	0.53
Deep Knockoff	[30]	0.74	1.00	0.29	0.88	0.25	0.95	0.40	0.86	0.26	0.94
DDLK	[38]	0.79	0.99	0.13	0.30	0.27	0.66	0.04	0.00	0.32	0.59
KnockoffGAN	[14]	0.44	0.99	0.07	0.35	0.09	0.70	0.05	0.17	0.09	0.60
sRMMD	[22]	0.72	1.00	0.29	0.88	0.24	0.94	0.31	0.78	0.26	0.93
DeepDRK	[34]	0.10	0.83	0.07	0.35	0.08	0.78	0.10	0.42	0.09	0.70
G^2M (ours)	–	**0.07**	**0.86**	0.09	0.58	**0.10**	**0.94**	0.10	0.32	**0.10**	**0.89**
G^2M† (ours)	–	**0.06**	**0.92**	**0.06**	**0.75**	**0.09**	**0.95**	**0.10**	**0.61**	**0.10**	**0.91**

For the Gaussian setup, we focus solely on the first set of methods, as they align with the theoretical guarantees for Gaussian designs. We use three different fitting methods to generate the estimation coefficient $\hat{\beta}_j$'s: ordinary least square (OLS), ridge regression, and LASSO, to consider common adaptation of these methods in the feature selection setup. In contrast, for the Copula and Gaussian mixture setups, we evaluate both sets of methods and consider ridge regression to be the fitting method as empirically it produces the best performance according to Shen et al. [34]. Results are presented for FDR and power with $(n, p) = (200, 100)$, at the FDR nominal level 0.1. All reported values are averaged over 100 independent repetitions.

Results: In Table 1, we present results with Gaussian data in two different correlation settings (e.g., $\rho = 0.6, 0.7$). In both cases, the proposed G^2M outperforms other benchmarking methods for achieving the highest power while controlling the FDR under the nominal level of 0.1. In addition, we notice that the power of the proposed G^2M is always greater than that of the Gaussian mirror and data splitting methods, indicating the consistency between the empirical results and the theoretical reasoning. Surprisingly, we discover that G^2M performs consistently across different fitting methods other than OLS (e.g., LASSO and ridge regression) despite the fact that the theory was developed in the least squares sense. This suggests a possible wide application of the proposed G^2M method.

In Table 2, on the other hand, we consider results with nonlinear design matrix X with ridge regression model being the fitting method. Similar to the Gaussian setting, G^2M achieves the highest power on all datasets while controlling the FDR under the nominal 0.1 level. The proposed method even outperforms those deep-learning-based methods, suggesting its power in wide applications. Importantly, this observation is consistent with the theoretical justification as the proof of the G^2M method does not depend on the distribution of the design matrix X.

To further demonstrate the performance of our method, we investigate high dimensional settings by reducing the number of samples relative to the 100 dimensional input data. In addition, we consider the impact of noise to the power and FDR during feature selection. The comparison is deferred to

[5] `https://github.com/msesia/deepknockoffs`

[6] `https://github.com/rajesh-lab/ddlk`

[7] `https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/knockoffgan`

[8] `https://github.com/ShoaibBinMasud/soft-rank-energy-and-applications`

Appendix E due to space limit. The results, however, suggest that our proposed method outperforms other benchmarking methods in most cases, indicating a new state-of-the-art.

4.2 Semi-synthetic Data

In this section, we conduct a semi-synthetic study by extracting the design matrix X from two real-world datasets.

Single-cell RNA Sequencing: The first dataset consists of single-cell RNA sequencing (scRNA-seq) data obtained from $10\times$ Genomics [9]. Each entry of $X \in \mathbb{R}^{n\times p}$ represents the observed expression levels of p genes across n cells. For additional background on this dataset, we refer readers to Hansen et al. [13] and Agarwal et al. [1]. Following the preprocessing pipeline described in Hansen et al. [13], we obtain the final dataset X with $n = 10,000$ and $p = 100$ [10]. The response y is formulated according to a nonlinear function of X. Details including the form of y and the data preparation about X are included in Appendix F due to space limit.

Table 3: FDR and power with two semi-synthetic datasets: RNA and IBD. **Bold** entries indicate the case with the highest power given controlled FDR at level 0.1, **Blue** for the second best, and Red for FDR>0.1.

Method	Citation	Semi-RNA FDR	Semi-RNA Power	Semi-IBD FDR	Semi-IBD Power
CRT	[8]	0.39	0.89	0.15	0.56
Distilled-CRT	[19]	0.27	0.87	0.11	0.75
Gaussian Mirror	[45]	0.61	0.97	0.04	0.26
Gaussian Mirror†	[45]	0.06	0.58	**0.08**	**0.55**
Data Splitting	[9]	0.21	0.86	0.16	0.73
Data Splitting†	[9]	0.14	0.64	0.16	0.77
HRT	[40]	0.38	0.87	0.05	0.42
Powerful Knockoff	[37]	0.50	0.94	0.07	0.21
Powerful Knockoff†	[37]	0.14	0.43	0.14	0.51
Deep Knockoff	[30]	0.00	0.14	0.27	0.55
DDLK	[38]	0.14	0.81	0.09	0.26
KnockoffGAN	[14]	0.00	0.00	0.10	0.25
sRMMD	[22]	0.00	0.00	0.24	0.44
DeepDRK	[34]	**0.08**	**0.73**	0.10	0.25
G^2M (ours)	–	0.00	0.43	0.07	0.45
G^2M† (ours)	–	**0.04**	**0.66**	**0.10**	**0.61**

Inflammatory Bowel Disease (IBD): The second dataset, publicly available from the Metabolomics Workbench [11], originates from a real-world study titled "Longitudinal Metabolomics of the Human Microbiome in Inflammatory Bowel Disease (IBD)" [20]. The objective of this study is to identify significant metabolites associated with two forms of inflammatory bowel disease: ulcerative colitis (UC) and Crohn's disease (CD). We use the C18 Reverse-Phase Negative Mode dataset, which comprises 546 samples and 91 metabolites.

To handle missing values, we preprocess the dataset by removing metabolites with more than 20% missing data, resulting in a final set of 80 metabolites. The data matrix is normalized entry-wise to have zero mean and unit variance, following a log transform and imputation of missing values using the k-nearest neighbor algorithm, as described in Masud et al. [22]. The response y uses the same linear model described in Section 4.1, which has lower signal strength compared to the experiment in Shen et al. [34].

In both semi-synthetic settings, we consider ridge regression as the fitting method as empirically it performs the best according to Shen et al. [34]. All reported values are averaged over 100 independent repetitions. The FDR nominal level is set to 0.1.

Results: Since we consider the nonlinear design matrix X, methods from both the deep-learning-based and non-deep-learning-based (see Section 4.1) are considered in this experiment. Results are presented in Table 3. We find that almost all non-deep-learning- and framework-based methods fail to control the FDR with the RNA data, and part of the methods fail on the IBD data. In comparison, the proposed G^2M method is the only one that can successfully control the FDR under the nominal level while achieving the second-best in power (the best-performing DeepDRK achieves a higher power of 0.73 in this case). In the IBD setting, on the other hand, G^2M beats all non-deep-learning-based methods and deep-learning-based methods for producing the highest power with the controlled FDR. This evidence suggests a stable application of G^2M on real-world data.

[9] `https://kb.10xgenomics.com/hc/en-us`

[10] Data processing code is available at: `https://github.com/dereklhansen/flowselect/tree/master/data`

[11] `https://www.metabolomicsworkbench.org/` under project DOI: 10.21228/M82T15

Table 4: Feature selection results with IBD real dataset that consider true X and y.: RNA and IBD. FDR level is specified as 0.2. Since there is no ground truth on the features, we report "number of referenced metabolites/number of identified" in place of FDR or power.

Model	G^2M (ours)†	CRT	Distilled-CRT	Gaussian Mirror†	Data Splitting†	HRT	Powerful Knockoff
Referenced / Identified	18/22	12/21	17/26	18/25	17/23	2/5	4/6
Model	DeepDRK	Deep Knockoff	sRMMD	KnockoffGAN	DDLK		
Referenced / Identified	19/23	15/20	5/5	12/14	17/25		

4.3 Real Case Study

We further conduct two case studies using real data (i.e., IBD dataset [20] and breast cancer dataset [43]) for both the design matrix X and the response variable y to qualitatively evaluate the selection performance of the proposed method. We defer the breast cancer dataset analysis in Appendix E.4.

In the IBD analysis, the response variable y is categorical, where $y = 1$ indicates that a sample is associated with UC/CD, and $y = 0$ otherwise. The covariates X are identical to those used in the second semi-synthetic setup described in Section 4.2. Since ground truth is not available for this dataset, we evaluate the results by identifying evidence of IBD-associated metabolites from existing literature, following the curation in Shen et al. [34], which draws upon the following sources: 1. Metabolites explicitly documented as being associated with IBD, UC, or CD in the PubChem database [12]; 2. Metabolites reported in peer-reviewed publications; 3. Metabolites mentioned in preprints. For convenience, we reproduce the nominal metabolite table from Shen et al. [34] in Appendix G.1.

We use the DeepPINK [21] model as the fitting method to obtain the feature coefficients in consideration of the nonlinearity between the input X and the response y in this real setting. The FDR nominal level is set to 0.2.

Results: We present feature selection results for the real IBD data in Table 4. Among the considered methods, clearly the G^2M method performs on par with the DeepDRK method and identifies more reported metabolites while keeping a lower number of total selections compared to other benchmarking methods. This reveals the potential of the G^2M method for real-world applications. For completeness, we include a full list of names for the identified metabolites by each method in Appendix G.2.

5 Conclusion

In this paper, we first identify a limitation of the existing mirror statistics in the data splitting paper—the strong unit variance assumption. We then proposed a variance-dependent Gaussian mirror method—G^2M—and show both theoretically and empirically the performance of G^2M compared to popular FDR-feature controlled frameworks and deep-learning-based knockoff methods with synthetic, semi-synthetic, and real datasets. The results demonstrate that the G^2M method effectively controls the FDR while achieving the highest power in most cases and delivering comparable performance in others. These findings suggest the potential for broad adoption of the G^2M method in real-world applications.

Limitations and broader impacts: one limitation of the work is the normal distribution assumption on the fitting coefficients (e.g., β_j^+ and β_j^-). This is tied to the proposed UMP test statistics in Lemma 3.5 and the fundamental result in Dai et al. [9]. A possible future work would be to generalize this part beyond the normality assumption in light of existing work on the generalized linear model setting: e.g., Dai et al. [10]. Nonetheless, based on experimental results with synthetic, semi-synthetic or real case study, we demonstrate that despite having this limitation in the theoretical formulation, our model still outperformed the existing state-of-the-art methods, suggesting the importance of the work and its potential use cases in biological data where dimensionality and FDR are crucial aspects.

[12] `https://pubchem.ncbi.nlm.nih.gov/`

References

[1] Divyansh Agarwal, Jingshu Wang, and Nancy R. Zhang. Data denoising and post-denoising corrections in single cell RNA sequencing. *Statistical Science*, 35(1):112 – 128, 2020.

[2] Ashwin N. Ananthakrishnan, Chengwei Luo, Vijay Yajnik, Hamed Khalili, John J. Garber, Betsy W. Stevens, Thomas Cleland, and Ramnik J. Xavier. Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell host & microbe*, 21(5):603–610, 2017.

[3] Armin Askari, Quentin Rebjock, Alexandre d'Aspremont, and Laurent El Ghaoui. Fanok: Knockoffs in linear time. *SIAM Journal on Mathematics of Data Science*, 3(3):833–853, 2021.

[4] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[5] Cristina Bauset, Laura Gisbert-Ferrándiz, and Jesús Cosín-Roger. Metabolomics as a promising resource identifying potential biomarkers for inflammatory bowel disease. *Journal of Clinical Medicine*, 10(4):622, 2021.

[6] Shoaib Bin Masud, Conor Jenkins, Erika Hussey, Seth Elkin-Frankston, Phillip Mach, Elizabeth Dhummakupt, and Shuchin Aeron. Utilizing machine learning with knockoff filtering to extract significant metabolites in Crohn's disease with a publicly available untargeted metabolomics dataset. *Plos one*, 16(7):e0255240, 2021.

[7] P. A. Blaker, M. Arenas-Hernandez, M. A. Smith, E. A. Shobowale-Bakre, L. Fairbanks, P. M. Irving, J. D. Sanderson, and A. M. Marinaki. Mechanism of allopurinol induced TPMT inhibition. *Biochemical pharmacology*, 86(4):539–547, 2013.

[8] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[9] Chenguang Dai, Buyu Lin, Xin Xing, and Jun S Liu. False discovery rate control via data splitting. *Journal of the American Statistical Association*, 118(544):2503–2520, 2023.

[10] Chenguang Dai, Buyu Lin, Xin Xing, and Jun S Liu. A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association*, 118 (543):1551–1565, 2023.

[11] DJ Fretland, DL Widomski, S Levin, and TS Gaginella. Colonic inflammation in the rabbit induced by phorbol-12-myristate-13-acetate. *Inflammation*, 14(2):143–150, 1990.

[12] Yeheng Ge, Sijia Zhang, and Xiao Zhang. False discovery rate control for high-dimensional cox model with uneven data splitting. *Journal of Statistical Computation and Simulation*, 94(7): 1462–1493, 2024.

[13] Derek Hansen, Brian Manzo, and Jeffrey Regier. Normalizing flows for knockoff-free controlled feature selection. In *Advances in Neural Information Processing Systems*, volume 35, pages 16125–16137, 2022.

[14] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. In *ICLR*, 2018.

[15] Hon Wai Koon. A novel orally active metabolite reverses Crohn's disease-associated intestinal fibrosis. *Inflammatory Bowel Diseases*, 28(Supplement_1):S61–S62, 2022.

[16] Aonghus Lavelle and Harry Sokol. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nature reviews Gastroenterology & hepatology*, 17(4):223–237, 2020.

[17] Thomas Lee, Thomas Clavel, Kirill Smirnov, Annemarie Schmidt, Ilias Lagkouvardos, Alesia Walker, Marianna Lucio, Bernhard Michalke, Philippe Schmitt-Kopplin, Richard Fedorak, et al. Oral versus intravenous iron replacement therapy distinctly alters the gut microbiota and metabolome in patients with IBD. *Gut*, 66(5):863–871, 2017.

[18] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[19] Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293, 2022.

[20] Jason Lloyd-Price, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, Nadim J. Ajami, Kevin S. Bonham, Colin J. Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019.

[21] Yang Lu, Yingying Fan, Jinchi Lv, and William Stafford Noble. DeepPINK: reproducible feature selection in deep neural networks. In *NeurIPS*, volume 31, 2018.

[22] Shoaib Bin Masud, Matthew Werenski, James M Murphy, and Shuchin Aeron. Multivariate soft rank via entropy-regularized optimal transport: Sample efficiency and generative modeling. *Journal of Machine Learning Research*, 24(160):1–65, 2023.

[23] Rishi S. Mehta, Zachary L. Taylor, Lisa J. Martin, Michael J. Rosen, and Laura B. Ramsey. *SLCO1B1* *15 allele is associated with methotrexate-induced nausea in pediatric patients with inflammatory bowel disease. *Clinical and translational science*, 15(1):63–69, 2022.

[24] Itta M. Minderhoud, Bas Oldenburg, Marguerite E. I. Schipper, Jose J. M. Ter Linde, and Melvin Samsom. Serotonin synthesis and uptake in symptomatic patients with Crohn's disease in remission. *Clinical Gastroenterology and Hepatology*, 5(6):714–720, 2007.

[25] Rajagopalan Lakshmi Narasimhan, Allison A. Throm, Jesvin Joy Koshy, Keith Metelo Raul Saldanha, Harikrishnan Chandranpillai, Rahul Deva Lal, Mausam Kumravat, Ajaya Kumar K. M., Aneesh Batra, Fei Zhong, et al. Inferring intestinal mucosal immune cell associated microbiome species and microbiota-derived metabolites in inflammatory bowel disease. *bioRxiv*, 2020.

[26] Binh T Nguyen, Bertrand Thirion, and Sylvain Arlot. A conditional randomization test for sparse logistic regression in high-dimension. *Advances in Neural Information Processing Systems*, 35:13691–13703, 2022.

[27] Andrea Nuzzo, Somdutta Saha, Ellen Berg, Channa Jayawickreme, Joel Tocker, and James R. Brown. Expanding the drug discovery space with predicted metabolite–target interactions. *Communications biology*, 4(1):1–11, 2021.

[28] Xiaofa Qin. Etiology of inflammatory bowel disease: a unified hypothesis. *World journal of gastroenterology: WJG*, 18(15):1708, 2012.

[29] Zhimei Ren and Rina Foygel Barber. Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):122–154, 2024.

[30] Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

[31] Sameh Saber, Rania M Khalil, Walied S Abdo, Doaa Nassif, and Eman El-Ahwany. Olmesartan ameliorates chemically-induced ulcerative colitis in rats via modulating NFκB and Nrf-2/HO-1 signaling crosstalk. *Toxicology and applied pharmacology*, 364:120–132, 2019.

[32] Thorsten Schmidt. Coping with copulas. *Copulas-From theory to application in finance*, 3: 1–34, 2007.

[33] Elizabeth A. Scoville, Margaret M. Allaman, Caroline T. Brown, Amy K. Motley, Sara N. Horst, Christopher S. Williams, Tatsuki Koyama, Zhiguo Zhao, Dawn W. Adams, Dawn B. Beaulieu, et al. Alterations in lipid, amino acid, and energy metabolism distinguish Crohn's disease from ulcerative colitis and control subjects by serum metabolomic profiling. *Metabolomics*, 14(1): 1–12, 2018.

[34] Hongyu Shen, Yici Yan, and Zhizhen Zhao. DeepDRK: Deep Dependency Regularized Knockoff for Feature Selection. In *Advances in Neural Information Processing Systems*, 2024.

[35] Johan D. Soderholm, Hans Oman, Lars Blomquist, Joggem Veen, Tuulikki Lindmark, and Gunnar Olaison. Reversible increase in tight junction permeability to macromolecules in rat ileal mucosa in vitro by sodium caprate, a constituent of milk fat. *Digestive diseases and sciences*, 43(7):1547–1552, 1998.

[36] Johan D. Söderholm, Gunnar Olaison, K. H. Peterson, L. E. Franzen, T. Lindmark, Mikael Wirén, Christer Tagesson, and Rune Sjödahl. Augmented increase in tight junction permeability by luminal stimuli in the non-inflamed ileum of Crohn's disease. *Gut*, 50(3):307–313, 2002.

[37] Asher Spector and Lucas Janson. Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, 50(1):252–276, 2022.

[38] Mukund Sudarshan, Wesley Tansey, and Rajesh Ranganath. Deep direct likelihood knockoffs. In *NeurIPS*, volume 33, 2020.

[39] Karsten Suhre, So-Youn Shin, Ann-Kristin Petersen, Robert P. Mohney, David Meredith, Brigitte Wägele, Elisabeth Altmaier, Panos Deloukas, Jeanette Erdmann, Elin Grundberg, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60, 2011.

[40] Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162, 2022.

[41] Kan Uchiyama, Shunichi Odahara, Makoto Nakamura, Shigeo Koido, Kiyohiko Katahira, Hiromi Shiraishi, Toshifumi Ohkusa, Kiyotaka Fujise, and Hisao Tajiri. The fatty acid profile of the erythrocyte membrane in initial-onset inflammatory bowel disease patients. *Digestive diseases and sciences*, 58(5):1235–1243, 2013.

[42] Victor Uko, Suraj Thangada, and Kadakkal Radhakrishnan. Liver disorders in inflammatory bowel disease. *Gastroenterology research and practice*, 2012(1):642923, 2012.

[43] W. Wolberg, O. Mangasarian, N. Street, and W. Street. Breast cancer wisconsin (diagnostic) [dataset]. UCI Machine Learning Repository, 1993. URL `https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic`.

[44] Xin Xing, Yu Gui, Chenguang Dai, and Jun S Liu. Neural gaussian mirror for controlled feature selection in neural networks. *arXiv preprint arXiv:2010.06175*, 2020.

[45] Xin Xing, Zhigen Zhao, and Jun S Liu. Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association*, 118(541):222–241, 2023.

[46] Yanjie Zhong, Todd Kuffner, and Soumendra Lahiri. Conditional randomization rank test. *arXiv preprint arXiv:2112.00258*, 2021.

Appendix: G^{2}M: A Generalized Gaussian Mirror Method to boost feature selection power

This appendix is structured as follows: in Appendix 2 we provide details on related work and the connection with our method; in Appendix A we show the empirical observation of the variance of Gaussian mirror statistics; in Appendix B we provide proofs for the theorems and lemmas in the main paper; in Appendix E we provide additional experimental results; in Appendix F we detail the RNA data preparation for experiment in Section 4.2; in Appendix G we provide additional information for the IBD study.

A Empirical Evidence on Variance Differences in Gaussian Mirror Statistics

This section provides empirical results about the variance (e.g. standard deviation) of the Gaussian mirror coefficients β_j^+ and β_j^- given different design matrices. To start, we randomly generate 10000 samples of the standard deviations of β_j^+ and β_j^- for both null and nonnull cases, respectively. Specifically, each sample is generated by first sampling a design matrix and then performing the Gaussian mirror perturbation based on a randomly sampled Gaussian vector z_j. Both the design matrix X and the sampled vector z_j are used to calculate the corresponding $c_j = \frac{\|P_{\perp -j} x_j\|}{\|P_{\perp -j} z_j\|}$—a component in the Gassuain mirror. The feature x_j is uniformly sampled on the index j given either null or nonnull. With the information of x_j, c_j and z_j, we are able to calculate the standard deviation of β_j^+ and β_j^- according to Corollary 3.4 and Eq. (2).

We present the histograms of the results with Gaussian design matrix in Figure 1 and 2, Gaussian mixture design matrix in Figure 3 and 4, and IBD design matrix in Figure 5 and 6. The description on the Gaussian, Gaussian mixture and IBD design matrices can be found in Section 4. Clearly, the standard deviations for both null and nonnull coefficients are distributions that are not concentrated at 1, indicating the unrealistic assumption about the unit variance outlined in Dai et al. [9].

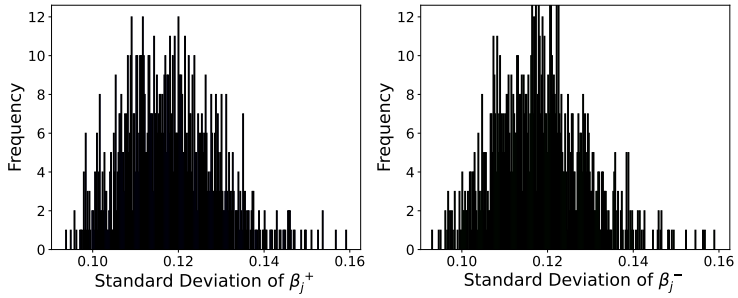

Figure 1: Histogram of the standard deviation of β_j^+ and β_j^- for null variables over 10000 samples. The design matrix X is based on Gaussian distributions.

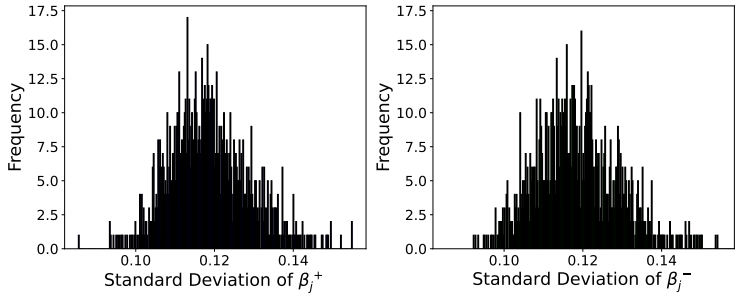

Figure 2: Histogram of the standard deviation of β_j^+ and β_j^- for nonnull variables over 10000 samples. The design matrix X is based on Gaussian distributions.

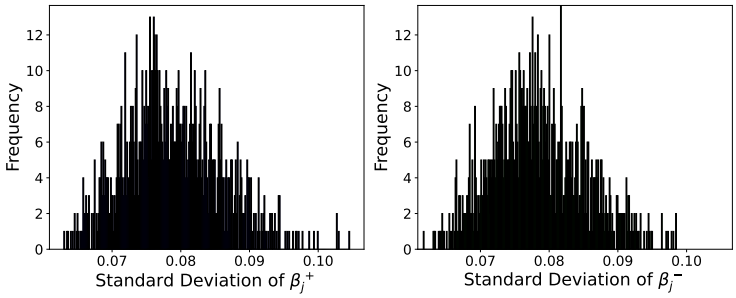

Figure 3: Histogram of the standard deviation of β_j^+ and β_j^- for null variables over 10000 samples. The design matrix X is based on Gaussian mixture distributions.

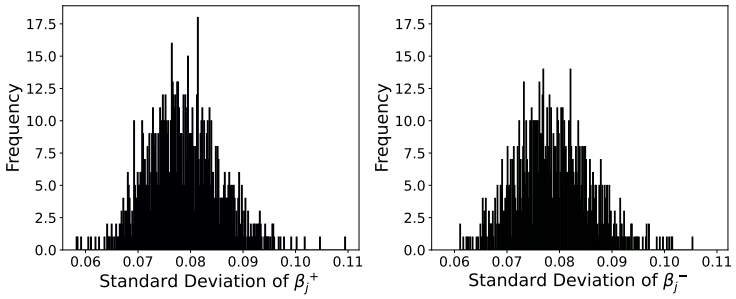

Figure 4: Histogram of the standard deviation of β_j^+ and β_j^- for nonnull variables over 10000 samples. The design matrix X is based on Gaussian mixture distributions.

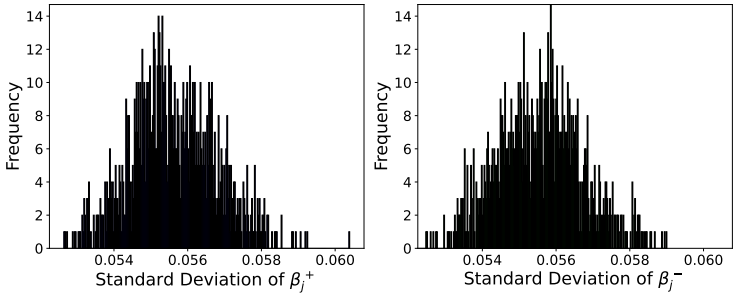

Figure 5: Histogram of the standard deviation of β_j^+ and β_j^- for null variables over 10000 samples. The design matrix X is based on the IBD dataset.

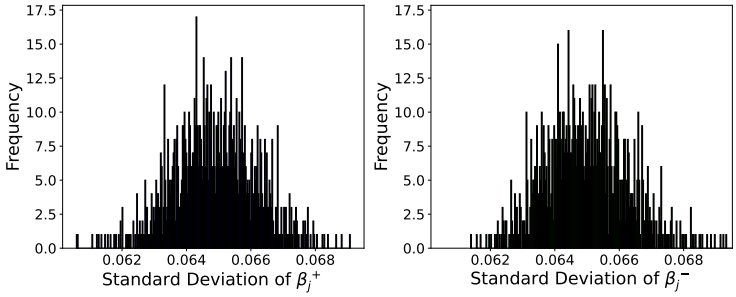

Figure 6: Histogram of the standard deviation of β_j^+ and β_j^- for nonnull variables over 10000 samples. The design matrix X is based on the IBD dataset.

B Proof

B.1 Proposition 3.1

To show proposition 3.1, we first consider two models. The original model that produces the response y and the new model that considers the inclusion of x_j^+ and x_j^- in place of the original x_j. In the following, we have the original model:

$$y = X\beta + \epsilon.$$

After modification of the design matrix, we have the new design matrix:

$$X_{\text{new}} = [X_{-j}, x_j^+, x_j^-],$$

and the new coefficient vector:

$$\beta_{\text{new}} = \begin{bmatrix} \beta_{-j}^{\text{new}} \\ \beta_j^+ \\ \beta_j^- . \end{bmatrix}$$

To perform the least square fit with the new design matrix and the original response variable y, we can expand normal equations $X_{\text{new}}^\top X_{\text{new}} \beta_{\text{new}} = X_{\text{new}}^\top y$:

$$\begin{bmatrix} X_{-j}^\top X_{-j} & X_{-j}^\top x_j^+ & X_{-j}^\top x_j^- \\ (x_j^+)^\top X_{-j} & (x_j^+)^\top x_j^+ & (x_j^+)^\top x_j^- \\ (x_j^-)^\top X_{-j} & (x_j^-)^\top x_j^+ & (x_j^-)^\top x_j^- \end{bmatrix} \begin{bmatrix} \beta_{-j}^{\text{new}} \\ \beta_j^+ \\ \beta_j^- \end{bmatrix} = \begin{bmatrix} X_{-j}^\top y \\ (x_j^+)^\top y \\ (x_j^-)^\top y \end{bmatrix}.$$

We use β_{-j}^{new} to distinguish from the original β_{-j} that represents the coefficients in the original linear model with j-th entry removed. Expanding the normal equation we have:

$$\begin{aligned}
(X_{-j}^\top x_j + c_j X_{-j}^\top z_j) \cdot \beta_j^+ + (X_{-j}^\top x_j - d_j X_{-j}^\top q_j) \cdot \beta_j^- &= X_{-j}^\top X_{-j} \cdot (\beta_{-j} - \beta_{-j}^{\text{new}}) + X_{-j}^\top x_j \cdot \beta_j + X_{-j}^\top \epsilon, \\
(x_j^{+\top} X_{-j}) \cdot \beta_{-j}^{\text{new}} + (x_j^{+\top} x_j^+) \cdot \beta_j^+ + (x_j^{+\top} x_j^-) \cdot \beta_j^- &= x_j^{+\top} X_{-j} \cdot \beta_j + x_j^{+\top} x_j \cdot \beta_j + x_j^{+\top} \epsilon, \\
(x_j^{-\top} X_{-j}) \cdot \beta_{-j}^{\text{new}} + (x_j^{-\top} x_j^+) \cdot \beta_j^+ + (x_j^{-\top} x_j^-) \cdot \beta_j^- &= x_j^{-\top} X_{-j} \cdot \beta_j + x_j^{-\top} x_j \cdot \beta_j + x_j^{-\top} \epsilon.
\end{aligned}$$

Since we have three equations and three unknowns: β_{-j}^{new}, β_j^+ and β_j^-, we cancel out β_{-j}^{new}, leaving two functions that are the functions of β_j^+ and β_j^-. Eventually, we obtain the final form that represents β_j^+ and β_j^- as a linear function of the true $beta_j$ and the noise ϵ:

$$\begin{aligned}
\beta_j^+ &= \alpha \cdot \beta_j + \gamma^\top \cdot \epsilon, \\
\beta_j^- &= \zeta \cdot \beta_j + \eta^\top \cdot \epsilon,
\end{aligned}$$

where

$$\alpha = \frac{N_{j^-} \cdot F - M_{j^-} \cdot H}{L}, \quad \gamma = \frac{N_{j^-} \cdot G - M_{j^-} \cdot K}{L},$$
$$\zeta = \frac{-N_{j^+} \cdot F + M_{j^+} \cdot H}{L}, \quad \eta = \frac{-N_{j^+} \cdot G + M_{j^+} \cdot K}{L}.$$

All the involved variables are presented below:

$$\begin{aligned}
L &= M_{j^+} \cdot N_{j^-} - M_{j^-} \cdot N_{j^+} \\
M_{j^+} &= \left(x_j^\top x_j + 2c z_j^\top x_j + c_j^2 z_j^\top z_j\right) - \left(x_j^\top X_{-j} + c_j z_j^\top X_{-j}\right) A^{-1} B, \\
M_{j^-} &= \left(x_j^\top x_j - d_j x_j^\top q_j + c_j z_j^\top x_j - c_j d_j z_j^\top q_j\right) - \left(x_j^\top X_{-j} + c_j z_j^\top X_{-j}\right) A^{-1} C, \\
N_{j^+} &= \left(x_j^\top x_j + c_j z_j^\top x_j - d_j q_j^\top x_j - c_j d_j q_j^\top z_j\right) - \left(x_j^\top X_{-j} - d_j q_j^\top X_{-j}\right) A^{-1} B, \\
N_{j^-} &= \left(x_j^\top x_j - 2d q_j^\top x_j + d_j^2 q_j^\top q_j\right) - \left(x_j^\top X_{-j} - d_j q_j^\top X_{-j}\right) A^{-1} C, \\
F &= (x_j + c_j z_j)^\top x_j - \left(x_j^\top X_{-j} + c_j z_j^\top X_{-j}\right) A^{-1} D, \\
G &= (x_j + c_j z_j)^\top \left(X_{-j} - c_j z_j^\top X_{-j}\right) A^{-1} X_{-j}^\top, \\
H &= (x_j - d_j q_j)^\top x_j - \left(x_j^\top X_{-j} - d_j q_j^\top X_{-j}\right) A^{-1} D, \\
K &= (x_j - d_j q_j)^\top \left(X_{-j} - d_j q_j^\top X_{-j}\right) A^{-1} X_{-j}^\top, \\
A &= X_{-j}^\top X_{-j}, \\
B &= X_{-j}^\top x_j + c_j X_{-j}^\top z_j, \\
C &= X_{-j}^\top x_j - d_j X_{-j}^\top q_j, \\
D &= X_{-j}^\top x_j, \\
E &= X_{-j}^\top \epsilon.
\end{aligned}$$

B.2 Corollary 3.2

To prove, we only need to show $L = M_{j^+} \cdot N_{j^-} - M_{j^-} \cdot N_{j^+}$ and $N_{j^-} \cdot F - M_{j^-} \cdot H + -N_{j^+} \cdot F + M_{j^+} \cdot H$ are the same. We first expand

$$M_{j^+} N_{j^-} - M_{j^-} N_{j^+} =$$

$$\begin{aligned}
&- 2cd\,(x_j^\top x_j)\,(q_j^\top x_j) - 2cd\,(z_j^\top x_j)\,(q_j^\top x_j) - 2d^2\,(x_j^\top q_j)^2 + d_j^2\,(q_j^\top q_j)\,(x_j^\top x_j) \\
&+ c_j^2\,(z_j^\top z_j)\,(x_j^\top x_j) - c_j^2\,(z_j^\top x_j)^2 - c_j^2 d_j\,(z_j^\top z_j)\,(x_j^\top q_j) + 2cd\,(z_j^\top q_j)\,(x_j^\top x_j) \\
&+ d_j^2 c_j\,(q_j^\top q_j)\,(z_j^\top x_j) + c_j^2 d_j\,(z_j^\top q_j)\,(z_j^\top x_j) - cd^2\,(z_j^\top q_j)\,(x_j^\top q_j) \\
&+ [\text{Terms related to } A^{-1} \text{and higher orders}].
\end{aligned}$$

Similarly, we expand

$$N_{j^-} \cdot F - M_{j^-} \cdot H + -N_{j^+} \cdot F + M_{j^+} \cdot H =$$

$$(M_{j^+} - M_{j^-})H + (N_{j^-} - N_{j^+})F =$$

$$\begin{aligned}
&c_j(z_j^\top x_j)(x_j^\top x_j) - cd(z_j^\top x_j)(x_j^\top q_j) + c_j^2(z_j^\top z_j)(x_j^\top x_j) - c_j^2 d_j(z_j^\top z_j)(x_j^\top q_j) \\
&+ d_j(x_j^\top q_j)(x_j^\top x_j) - d_j^2(x_j^\top q_j)^2 + cd(z_j^\top q_j)(x_j^\top x_j) - cd^2(z_j^\top q_j)(x_j^\top q_j) \\
&- d_j(q_j^\top x_j)(x_j^\top x_j) - dc(q_j^\top x_j)(z_j^\top x_j) + cd(z_j^\top q_j)(x_j^\top x_j) + d_j^2(q_j^\top q_j)(x_j^\top x_j) \\
&+ d_j^2 c_j(q_j^\top q_j)(z_j^\top x_j) - c_j^2(z_j^\top x_j)^2 - c_j(z_j^\top x_j)(x_j^\top x_j) + c_j^2 d_j(z_j^\top q_j)(z_j^\top x_j) \\
&[\text{Terms involving } A^{-1} \text{and higher orders}].
\end{aligned}$$

We can find matches for every term, proving the equality.

B.3 Corollary 3.3

The proof is similar to the one in Corollary 3.2, we expand terms and show the match. In this proof, we need to show $N_{j^-} \cdot F - M_{j^-} \cdot H = -N_{j^+} \cdot F + M_{j^+} \cdot H$. We expand both sides and compare, which proves the corollary. In the following, we have:

$$-N_{j^+} \cdot F + M_{j^+} \cdot H =$$

$$\begin{aligned}
&c_j^2\Big[(z_j^\top z_j)(x_j^\top x_j) - (z_j^\top x_j)^2\Big] \\
&- cd(q_j^\top x_j)(z_j^\top x_j) + cd(q_j^\top z_j)(x_j^\top x_j) \\
&+ c_j^2 d_j(q_j^\top z_j)(z_j^\top x_j) - c_j^2 d_j(z_j^\top z_j)(q_j^\top x_j) \\
&+ [\text{Complex terms involving } X_{-j}, A^{-1}],
\end{aligned}$$

$$N_{j^-} \cdot F - M_{j^-} \cdot H =$$

$$\begin{aligned}
&\Big[- cd(q_j^\top x_j)(z_j^\top x_j) + cd(z_j^\top q_j)(x_j^\top x_j)\Big] \\
&+ d_j^2\Big[(q_j^\top q_j)(x_j^\top x_j) - (q_j^\top x_j)^2\Big] \\
&+ cd^2\Big[(q_j^\top q_j)(z_j^\top x_j) - (q_j^\top x_j)(z_j^\top q_j)\Big] \\
&+ \big[\text{Complex terms involving } X_{-j}, A^{-1}\big].
\end{aligned}$$

After considering $c_j = d_j$ and $z_j = q_j$, we can verify the two functions are equal, completing the proof.

Remark B.1. Note that the Gaussian mirror work [45] only presents the $\beta_j^+ = \beta_j^- = 0.5$ without providing the formal proof. Combining Corollary 3.2 and Corollary 3.3, we can provide formal evidence for this statement.

B.4 Lemma 3.5

This proof is inspired by [9] for a more general setting. First, we identify the general form of the statistics following the proof in [9]. We then provide explicitly the form of f (e.g., in Eq. (3)) according to the Neyman-Pearson lemma. The proof of the latter is omitted in [9]. We hope this proof can serve as a complement, which provides insights for people who are interested in any future extension on the optimal test statistics in this more general setting.

To start with, let $Z_1, Z_2 \sim N(\omega, \sigma_1^2), N(\omega, \sigma_2^2)$, respectively. $Z_3, Z_4 \sim N(0, \sigma_3^2)$ and $N(0, \sigma_4^2)$, respectively [13]. And $\omega \sim \delta \cdot \text{Rademacher}(0.5)$, $\delta > 0$. All variables are independent. Following [9], we assume that the designated FDR control level $q \in (0, 1)$ satisfies $\frac{rq}{1-q} < 1$, otherwise selecting all features would maximize the power and also achieve asymptotic FDR control. Let $f_{\text{opt}}(u, v)$ be the optimal choice, and let S_{opt} be the optimal selection result that achieves asymptotic FDR control. By the law of large numbers, we have:

$$\lim_{p\to\infty} \frac{\#\{j : j \in S_0, j \in S_{\text{opt}}\}}{\#\{j : j \in S_{\text{opt}}\}} = \frac{P(j \in S_{\text{opt}} | j \in S_0)}{P(j \in S_{\text{opt}} | j \in S_0) + rP(j \in S_{\text{opt}} | j \in S_1)} \leq q, \tag{5}$$

in which the numerator is the type-I error. More precisely:

$$P(j \in S_{\text{opt}} | j \in S_1) = P(\text{sign}(Z_1 Z_2) f_{\text{opt}}(|Z_1|, |Z_2|) > t_{\text{opt}}).$$
$$P(j \in S_{\text{opt}} | j \in S_0) = P(\text{sign}(Z_3 Z_4) f_{\text{opt}}(|Z_3|, |Z_4|) > t_{\text{opt}}),$$

Here, $t_{\text{opt}} > 0$ is the cutoff that maximizes the power $P(j \in S_{\text{opt}} | j \in S_1)$, under the constraint that Eq. (5) holds.

In practice, we consider testing whether the covariate X_j is a null feature, with the significance level α specified as:

$$\alpha = \frac{rq}{1-q} P(j \in S_{\text{opt}} | j \in S_1) < 1.$$

Given two realizations β_j^+ and β_j^-, we consider their random variables independently following the normal distributions described by Z_1, Z_2 if X_j is a nonnull feature, and by Z_3, Z_4 otherwise. According to Eq. (5), the test which rejects the null hypothesis (i.e., $j \in S_{\text{opt}}$) if:

$$\text{sign}(\beta_j^+ \beta_j^-) f_{\text{opt}}(|\beta_j^+|, |\beta_j^-|) > t_{\text{opt}} \tag{6}$$

achieves the significance level α.

We consider the following rejection rule with a certain form of f that will be revealed later in the proof:

$$\text{sign}(\beta_j^+ \beta_j^-) f(|\beta_j^+|, |\beta_j^-|) > t_{\text{lik}},$$

in which $t_{\text{lik}} > 0$ satisfies:

$$P(f(|Z_3|, |Z_4|) > t_{\text{lik}} | \text{sign}(Z_3) = \text{sign}(Z_4)) = 2\alpha.$$

Let S_{lik} be the corresponding selection set. We first show that this rejection rule controls the type-I error below α. Indeed:

$$P(j \in S_{\text{lik}} | j \in S_0) = \frac{1}{2} P(f(|\beta_j^+|, |\beta_j^-|) > t_{\text{lik}} | j \in S_0, \text{sign}(\beta_j^+) = \text{sign}(\beta_j^-)) = \alpha.$$

In terms of power, we have:

$$P(j \in S_{\text{lik}} | j \in S_1) = p_w P(f(|\beta_j^+|, |\beta_j^-|) > t_{\text{lik}} | j \in S_1, \text{sign}(\beta_j^+) = \text{sign}(\beta_j^-))$$
$$\geq p_w P(f_{\text{opt}}(|\beta_j^+|, |\beta_j^-|) > t_{\text{opt}} | j \in S_1, \text{sign}(\beta_j^+) = \text{sign}(\beta_j^-)) = P(j \in S_{\text{opt}} | j \in S_1),$$

where $p_w = P\left(\text{sign}(\beta_j^{(1)}) = \text{sign}(\beta_j^{(2)}) \mid j \in S_1\right)$.

In the following, we show the proof of the inequality and the exact form of the function f via the Neyman-Pearson lemma.

[13] Note that we consider a more general setting with σ_1, σ_2, σ_3, and σ_4. Later on, we assume $\sigma_1 = \sigma_3 = \sigma_a$ and $\sigma_2 = \sigma_4 = \sigma_b$, to reach the results in Lemma 3.5.

Since we consider a function f that takes the realizations as inputs (e.g., Eq. (6)), we first need to study the distribution of $|\beta_j^+|, |\beta_j^-|$ conditioning on the equal sign constraint: $\text{sign}(\beta_j^+) = \text{sign}(\beta_j^-)$.

We first define the event E to represent $\text{sign}(\beta_j^+) = \text{sign}(\beta_j^-)$. Then the events would have two cases: 1. both terms are positive; 2. both terms are negative. Under H_0, since β_j^+ and β_j^- are independent and symmetric around zero. Therefore:

$$P_{E|H_0} = P_{H_0}(\beta_j^+ > 0, \beta_j^- > 0) + P_{H_0}(\beta_j^+ < 0, \beta_j^- < 0) = 0.25 + 0.25 = 0.5.$$

Under H_1, on the other hand, we have:

$$P_{E|H_1} = 0.5 \cdot P_{E|w=\delta} + 0.5 \cdot P_{E|w=-\delta},$$

where

$$P_{E|w=\delta} = P_{E|w=-\delta} = \Phi\left(\frac{\delta}{\sigma_1}\right)\Phi\left(\frac{\delta}{\sigma_2}\right) + \left[1 - \Phi\left(\frac{\delta}{\sigma_1}\right)\right]\left[1 - \Phi\left(\frac{\delta}{\sigma_2}\right)\right],$$

with Φ being the cumulative density function of the standard normal distribution. Combining the two equations above, we obtain:

$$P_{E|H_1} = \Phi\left(\frac{\delta}{\sigma_1}\right)\Phi\left(\frac{\delta}{\sigma_2}\right) + \left[1 - \Phi\left(\frac{\delta}{\sigma_1}\right)\right]\left[1 - \Phi\left(\frac{\delta}{\sigma_2}\right)\right].$$

After characterizing the probability of the equal sign event, we consider two conditional density functions under H_0 and H_1, respectively. Specifically under H_0:

$$f_{|\beta_j^+|,|\beta_j^-||E,H_0}(a,b) = \frac{P_{H_0}(\beta_j^+ > 0, \beta_j^- > 0) \cdot f_{H_0}(a,b) + P_{H_0}(\beta_j^+ < 0, \beta_j^- < 0) \cdot f_{H_0}(-a,-b)}{P_{E|H_0}},$$

where $f_{H_0}(\beta_j^+, \beta_j^-) = f_{\beta_j^+}(\beta_j^+; 0, \sigma_3^2) \cdot f_{\beta_j^-}(\beta_j^-; 0, \sigma_4^2)$.

Similarly under H_1, we have:

$$f_{|\beta_j^+|,|\beta_j^-||E,H_1}(a,b) = \frac{P_{++} \cdot f_{++}(a,b) + P_{--} \cdot f_{--}(a,b)}{P_{E|H_1}},$$

where:

$$f_{++}(a,b) = f_{\beta_j^+}(a; \delta, \sigma_1^2) \cdot f_{\beta_j^-}(b; \delta, \sigma_2^2),$$
$$f_{--}(a,b) = f_{\beta_j^+}(a; -\delta, \sigma_1^2) \cdot f_{\beta_j^-}(b; -\delta, \sigma_2^2),$$
$$P_{++} = P_{--} = \frac{1}{2}\left[\Phi\left(\frac{\delta}{\sigma_1}\right)\Phi\left(\frac{\delta}{\sigma_2}\right) + \Phi\left(-\frac{\delta}{\sigma_1}\right)\Phi\left(-\frac{\delta}{\sigma_2}\right)\right].$$

Overall, the likelihood ratio between H_1 and H_0 can be represented as:

$$\Lambda(a,b) = \frac{f_{|\beta_j^+|,|\beta_j^-||E,H_1}(a,b)}{f_{|\beta_j^+|,|\beta_j^-||E,H_0}(a,b)}$$
$$= U \cdot [P_+ \exp(-S_-) + P_- \exp(-S_+)]$$

where

$$U = \left(\frac{1}{P_{E|H_1}}\right)\left(\frac{\sigma_3\sigma_4}{\sigma_1\sigma_2}\right),$$
$$S_- = \frac{(a-\delta)^2}{2\sigma_1^2} + \frac{(b-\delta)^2}{2\sigma_2^2} - \left[\frac{a^2}{2\sigma_3^2} + \frac{b^2}{2\sigma_4^2}\right],$$
$$S_+ = \frac{(a+\delta)^2}{2\sigma_1^2} + \frac{(b+\delta)^2}{2\sigma_2^2} - \left[\frac{a^2}{2\sigma_3^2} + \frac{b^2}{2\sigma_4^2}\right],$$
$$P_+ = \Phi\left(\frac{\delta}{\sigma_1}\right)\Phi\left(\frac{\delta}{\sigma_2}\right),$$
$$P_- = \left[1 - \Phi\left(\frac{\delta}{\sigma_1}\right)\right]\left[1 - \Phi\left(\frac{\delta}{\sigma_2}\right)\right].$$

In addition, we let $\sigma_1 = \sigma_3 = \sigma_a$ and $\sigma_2 = \sigma_4 = \sigma_b$, leading to:

$$U = \left(\frac{1}{P_{E|H_1}}\right),$$

$$S_- = \frac{\delta(\delta - 2a)}{2\sigma_a^2} + \frac{\delta(\delta - 2b)}{2\sigma_b^2},$$

$$S_+ = \frac{\delta(\delta + 2a)}{2\sigma_a^2} + \frac{\delta(\delta + 2b)}{2\sigma_b^2},$$

$$P_+ = \Phi\left(\frac{\delta}{\sigma_a}\right)\Phi\left(\frac{\delta}{\sigma_b}\right),$$

$$P_- = \left[1 - \Phi\left(\frac{\delta}{\sigma_a}\right)\right]\left[1 - \Phi\left(\frac{\delta}{\sigma_b}\right)\right].$$

Letting $\Lambda(a, b)$ be $f(a, b)$ completes the proof, as $\Lambda(a, b)$ is the UMP test statistic according to the Neyman-Pearson lemma.

B.5 Theorem 3.6

To show that the proposed test statistic $\text{sign}(ab)f(a, b)$ (Eq. (3)) is better than the Gaussian mirror [45] and data splitting [9] counterparts, we prove with two parts. In the first part, we show that the mirror statistics can vary for different j's. Secondly, we show that the proposed test statistic (Eq. (3)) is UMP given j.

B.5.1 Part I

Differing from existing works like Gaussian mirror or data splitting that considers a generic distribution for all β_j under H_0 and H_1, we interpret the FDR under H_0 from another angle that considers the changes in distribution to β_j across different j's. Namely, given an arbitrary mirror test statistic $\text{sign}(\beta_j^+\beta_j^-)f_j(\beta_j^+, \beta_j^-)$ that depends on the index j, and a general index-agnostic counterpart $\text{sign}(\beta_j^+\beta_j^-)f(\beta_j^+, \beta_j^-)$, the FDR can be defined as:

$$P(\text{sign}(\beta^+\beta^-)f(\beta^+, \beta^-) < t_q|H_0) = \sum_{j\in[1,\ldots,p]} P(j)P(\text{sign}(\beta_j^+\beta_j^-)f_j(\beta_j^+, \beta_j^-) < t_q^j|H_0, j), \tag{7}$$

where $P(j)$ refers to the probability of the presence of the index j. In practice, we do not need to know the specification of this distribution as the proof relies only on the symmetry of the conditional distribution: $P(\text{sign}(\beta_j^+\beta_j^-)f_j(\beta_j^+, \beta_j^-)|H_0, j)$. Apparently $\sum_{j\in[1,\ldots,p]} P(j) = 1$ holds. t_q is a threshold chosen such that the FDR can be controlled by the nominal level q. t_q^j adds the dependence on the index j. In previous works, it is common to assume identical β_j in H_0 and H_1 to have the same distribution across all j's, leading to the fact that $\text{sign}(\beta_j^+\beta_j^-)f(\beta_j^+, \beta_j^-) < t_q$ is the same as $\text{sign}(\beta^+\beta^-)f(\beta^+, \beta^-) < t_q$. However, if one considers the selection rule for these approaches (e.g., Gaussian mirror/data splitting), it is clear that such a universal assumption need not to held. This is because to control the FDR, the only necessary property is the symmetry of the null distribution. According to Eq. 7, we realize that even in the case where $\text{sign}(\beta_j^+\beta_j^-)f(\beta_j^+, \beta_j^-)$ varies across j's, the overall FDR can still be controlled if all of these distributions are symmetric about zero. The only difference is that for each $\text{sign}(\beta_j^+\beta_j^-)f(\beta_j^+, \beta_j^-)$, the threshold t_q^j is different given different distributions. And the proof of higher power is straightforward given this distributional flexibility.

B.5.2 Part II

According to Lemma 3.5, we already showed that the proposed test statistic $\text{sign}(\beta_j^+\beta_j^-)f(\beta_j^+, \beta_j^-)$ in Eq. (3) is UMP given the setting considered in this paper. This means that given any arbitrary t_q^j for some FDR level (not necessarily q as q refers to the general selection rule across all β_j's), we can achieve the highest power compared to the test statistics in Gaussian mirror and data splitting. This leads to the conclusion that the overall power is the highest, hence completing the proof.

Remark B.2. This proof simply reveals a fact that under the controlled nominal FDR level q, to find a better test statistic, we only need to maximize the power of individual test statistic for every β_j (or, equivalently, X_j), rather than a general form of test statistic across all $j \in [1, \ldots, p]$.

To the best of our knowledge, we are the first to provide this reasoning in the proof and we hope this brings insights into developing the UMP test statistics, which is beyond the scope of this paper.

C Complexity Discussion between Benchmarking Methods

The complexity of the Gaussian Mirror method is $\mathcal{O}(np^3 + p^4)$ (for $n > p$). Essentially it runs p ordinary least square fit (OLS), each of which has $\mathcal{O}(np^2 + p^3)$ complexity ($n > p$). The computational complexity of Algorithm 2 is $\mathcal{O}(np^3 + p^4 + pki)$ (for $n > p$), where the $\mathcal{O}(pki)$ part is introduced by the k-means (following Lloyd's algorithm). k is the number of clusters and i stands for the number of iterations until convergence. In comparison, the data splitting runs two OLS, resulting in $\mathcal{O}(np^2 + p^3)$ complexity. The knockoff framework, according to Askari et al. [3], requires at least $\mathcal{O}(p^{3.5})$ to solve for the knockoff variable in a semi-definite programming setting. Later on, it requires one OLS for $2p$ dimension, resulting in $\mathcal{O}(p^{3.5} + 4np^2 + 8p^3)$ complexity. Other deep-learning-based knockoff variables require additional deep learning model fitting to obtain the knockoff statistics, making the complexity analysis hard given the choice of optimizer and the model architecture, hence ignored. According to Zhong et al. [46], the computational complexity for CRT, assuming $p = n$ is $\mathcal{O}(p^3 \log^2 p)$. We did not find any rigorous complexity analysis about HRT and dCRT, however, based on the results reported in Table 2 of the paper [26], we believe the complexity is between model-X knockoff and CRT.

D Design Matrix Setup for Synthetic Experiment

This section presents the model applied to the design matrix X in the synthetic dataset setup in Section 4.1.

Gaussian: We replicate the multivariate normal benchmark described in Romano et al. [30]. Specifically, we sample $x \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where Σ is a d-dimensional covariance matrix with entries $\Sigma_{i,j} = \rho^{|i-j|}$. This autoregressive Gaussian data structure captures strong correlations between adjacent features, with diminishing correlations as the distance between features increases. For our experiments, we consider $\rho \in \{0.5, 0.6, 0.7\}$ to provide additional context on how the change of ρ affects the feature selection performance, in addition to the chosen 0.6 from Romano et al. [30]. Due to limited space in the main paper, we present results for $\rho = 0.5$ in Appendix E.1.

Gaussian mixture: We utilize a Gaussian mixture model for X, represented as $X \sim \sum_{k=1}^{3} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$, where the proportions of the three Gaussian components are given by $(\pi_1, \pi_2, \pi_3) = (0.4, 0.2, 0.4)$. The mean vectors $\mu_k \in \mathbb{R}^p$ are defined as $\mu_k = \mathbf{1}_p \cdot 20 \cdot (k-1)$, where $\mathbf{1}_p$ is a p-dimensional vector of ones. The covariance matrices $\Sigma_k \in \mathbb{R}^{p \times p}$ have entries (i, j) computed as $\rho_k^{|i-j|}$, where $\rho_k = \rho_{\text{base}}^{k-0.1}$ and $\rho_{\text{base}} = 0.6$. Both **Gaussian** and **Gaussian mixture** implementation can be found in [34] [14].

Copula Models: To evaluate data structures with more complex correlation, we incorporate copula models [32]. Specifically, we consider two copula families: Clayton and Joe, each parameterized with a consistent copula parameter of 2. Marginal distributions are selected as either a uniform distribution (via identity transformation) or an exponential distribution with a rate of 1. These copulas are implemented using the `PyCop` library [15]. Essentially, copulas are statistical tools designed to model and simulate complex dependencies among random variables, independently from the shapes of their marginal distributions. Unlike traditional multivariate models that assume linear or Gaussian relationships, copulas allow us to construct datasets where variables exhibit non-linear or asymmetric dependencies, better reflecting patterns seen in real-world data. In our study, we use two widely-studied copula families: the Clayton copula, which models strong lower-tail dependence (i.e., variables tend to move together when their values are low), and the Joe copula, which captures strong upper-tail dependence (i.e., variables tend to move together when their values are high). By

[14] `https://github.com/nowonder2000/DeepDRK`

[15] `https://github.com/maximenc/pycop/`

specifying a copula parameter of 2, we control the strength of these dependencies in a consistent way across scenarios. For each simulated dataset, the individual variable distributions (marginals) are chosen to be either uniform (via identity transformation) or exponential (rate=1), allowing us to assess the robustness of our methods under different data distributions. This setup enables a comprehensive evaluation of how our proposed methods perform under diverse and realistic correlation structures.

E Additional Experiment

This section extends the experiments in the main paper to cover e.g., the high dimensional setting and various noise levels, to further demonstrate the performance of the proposed method. In particular, we consider synthetic experiments and focus on the ridge regression model for the fitting coefficients (see Sec. 4.1) for details).

E.1 Additional Gaussian Synthetic Dataset Result

This section presents benchmarking results for the Gaussian setting complementing Table 1 with $rho = 0.5$. Results are presented in Table 5.

Table 5: FDR and power with Gaussian design matrix X for $\rho = 0.5$. **Bold** entries indicate the case with the highest power given controlled FDR level: 0.1. **Blue** entries indicate the second best. Red entries indicate FDR> 0.1.

Method	OLS		Ridge		LASSO	
	FDR	Power	FDR	Power	FDR	Power
CRT [8]	0.38	0.93	0.27	0.95	0.17	0.99
Distilled-CRT [19]	0.54	0.99	0.34	0.99	0.15	0.98
Data Splitting [9]	0.00	0.00	0.13	0.82	0.08	0.70
Data Splitting† [9]	0.06	0.75	0.13	0.82	0.13	0.76
Gaussian Mirror [45]	0.10	0.79	0.05	0.77	0.07	0.84
Gaussian Mirror† [45]	**0.10**	**0.82**	**0.05**	**0.80**	**0.07**	**0.87**
HRT [40]	0.00	0.34	0.00	0.36	0.01	0.36
Powerful Knockoff [37]	0.00	0.00	0.07	0.49	0.05	0.77
Powerful Knockoff† [37]	0.06	0.80	0.09	0.60	0.05	0.80
G^2M (ours)	**0.10**	**0.88**	**0.07**	**0.90**	**0.07**	**0.94**
G^2M† (ours)	**0.10**	**0.88**	**0.07**	**0.90**	**0.07**	**0.94**

E.2 High Dimensional Setting

We extend Table 1 and Table 2 with two additional low sample-size settings (i.e., $n/p = 30/100$ and $90/100$), considering ridge regression for its better performance compared to LASSO.

Notably, given that we already consider a low signal strength setting in both synthetic and semi-synthetic experiments, following the experimental setup in Shen et al. [34] (see Sec. 4 in the main paper), to balance the reduced sample size with signal strength, we apply a multiplier to boost the signal strength. This is a common practice in statistical analysis. Specifically, we use the multiplier $\sqrt{200}/\sqrt{p}$, where $p = 30$ or 100, to ensure a fair comparison between the new cases and the $n/p = 200/100$ case in Sec. 4.

Additionally, for Data Splitting, Gaussian Mirror, and the proposed G^2M, we consider the non-dagger versions, as we observed that the dagger versions cannot properly control FDR when $p > n$. The results for the Gaussian setting and the non-Gaussian setting are included in Table 6 and 7, respectively.

From Table 6, we observe that the proposed G^2M outperforms all other methods by achieving the highest power with controlled FDR. In addition, Table 7 suggests that G^2M is generally a good feature selection algorithm for having the best or the second best power with controlled FDR. More importantly, in this case, G^2M never exceeds the FDR nominal level of 0.1, suggesting its soundness in the proposed theoretical framework.

Table 6: Extension of Table 1 with high-dimensional settings: FDR / power under two different n/p ratios and two correlation strengths. Red entries indicate FDR failures (> 0.1), **Bold** entries indicate the highest power among methods controlling FDR ≤ 0.1, and **Blue** entries indicate the second highest.

Method	$n/p = 90/100$		$n/p = 30/100$	
	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.6$	$\rho = 0.7$
CRT [8]	0.00 / 0.15	0.00 / 0.09	0.00 / 0.00	0.00 / 0.01
Distilled-CRT [19]	0.59 / 0.95	0.59 / 0.92	**0.04 / 0.06**	0.09 / 0.09
Gaussian Mirror [45]	0.11 / 0.55	**0.08 / 0.54**	0.01 / 0.03	0.00 / 0.00
Data Splitting [9]	**0.08 / 0.53**	0.07 / 0.52	0.02 / 0.05	**0.06 / 0.10**
HRT [40]	0.00 / 0.08	0.00 / 0.03	0.00 / 0.00	0.00 / 0.00
Powerful Knockoff [37]	0.02 / 0.11	0.03 / 0.13	0.00 / 0.00	0.00 / 0.00
G^2M (ours)	**0.06 / 0.58**	**0.08 / 0.55**	**0.09 / 0.13**	**0.05 / 0.11**

E.3 Noise Level Variation

To study robustness to noise, we follow Shen et al. [34] and vary the signal strength by scaling β as $\frac{p}{10\sqrt{n}}$, $\frac{p}{15\sqrt{n}}$ (already in the paper), and $\frac{p}{20\sqrt{n}}$. Below, we report results for the additional $\frac{p}{10\sqrt{n}}$ and $\frac{p}{20\sqrt{n}}$ settings. All other experimental parameters remain the same. Similarly, ridge regression is considered for all methods. And results for the Gaussian setting and the non-Gaussian setting are included in Table 8 and 9, respectively.

From Table 8, we observe that G^2M beats other methods by having the highest power with controlled FDR—a case that is similarly revealed in Table 6. Besides, Table 9 also suggests a good performance of the proposed G^2M for having 8 out of 10 best performing results (i.e., consider the highest power with controlled FDR). This indicate the outperformance of the proposed G^2M in both Gaussian and non-Gaussian settings.

E.4 Another Real Case Study—Breast Cancer Dataset

We conduct an additional real-world case study using the Breast Cancer Wisconsin (Diagnostic) dataset [43], which consists of 569 patient samples. Each sample is labeled with a binary diagnostic outcome (malignant vs. benign), accompanied by 30 quantitative features extracted from digitized images of fine-needle aspirate (FNA) of breast mass tissue. Following the same evaluation protocol as our previous case study, we reviewed the biomedical literature to identify features that have been consistently reported as highly indicative of malignant tissue. These literature-supported features serve as the ground truth for our analysis. Based on this reference set, we report in Table X the number of "referenced vs. identified" features selected by each competing method. In total, 22 features are recognized as clinically relevant. Results are in Table 10.

F Preparation of the RNA Data

We first normalize the raw data X to value range $[0, 1]$ and then standardize it to have zero mean and unit variance. Y is synthesized according to X. We consider the response Y is generated following the expression:

$$\begin{aligned} k &\in [m/4] \\ \varphi_k^{(1)}, \varphi_k^{(2)} &\sim \mathcal{N}(1,1) \\ \varphi_k^{(3)}, \varphi_k^{(4)}, \varphi_k^{(5)} &\sim \mathcal{N}(2,1) \\ Y \mid X = \epsilon + \sum_{k=1}^{m/4} & \varphi_k^{(1)} X_{4k-3} + \varphi_k^{(3)} X_{4k-2} \\ + \varphi_k^{(4)} \tanh & \left(\varphi_k^{(2)} X_{4k-1} + \varphi_k^{(5)} X_{4k} \right), \end{aligned} \tag{8}$$

Table 7: Extension of Table 2 with high-dimensional settings: FDR and power for each dataset under two sample-size regimes. Red entries indicate FDR>0.1, **Bold** for the highest power among methods controlling FDR ≤ 0.1, and **Blue** for the second highest.

Dataset	Method	FDR / Power $n/p = 90/100$	FDR / Power $n/p = 30/100$
GaussianMixtureAR1	CRT [8]	0.00 / 0.00	0.00 / 0.00
	Distilled-CRT [19]	0.10 / 0.13	0.00 / 0.00
	Gaussian Mirror [45]	**0.06 / 0.52**	0.01 / 0.01
	Data Splitting [9]	0.05 / 0.32	**0.06 / 0.07**
	HRT [40]	0.00 / 0.01	0.00 / 0.01
	Deep Knockoff [30]	0.77 / 1.00	0.79 / 0.97
	DDLK [38]	0.25 / 0.21	0.00 / 0.00
	KnockoffGAN [14]	0.02 / 0.10	0.00 / 0.01
	sRMMD [22]	0.75 / 1.00	0.79 / 0.97
	DeepDRK [34]	0.01 / 0.20	0.00 / 0.00
	G^2M (ours)	**0.04 / 0.39**	**0.04 / 0.04**
Copula: Clayton & Exp.	CRT [8]	0.00 / 0.00	0.00 / 0.00
	Distilled-CRT [19]	0.12 / 0.21	**0.08 / 0.04**
	Gaussian Mirror [45]	0.03 / 0.10	0.11 / 0.14
	Data Splitting [9]	**0.07 / 0.20**	0.05 / 0.02
	HRT [40]	0.00 / 0.00	0.00 / 0.00
	Deep Knockoff [30]	0.47 / 0.88	0.77 / 0.84
	DDLK [38]	0.00 / 0.00	0.00 / 0.00
	KnockoffGAN [14]	0.00 / 0.00	0.00 / 0.00
	sRMMD [22]	0.42 / 0.81	0.75 / 0.81
	DeepDRK [34]	0.08 / 0.15	0.17 / 0.35
	G^2M (ours)	**0.05 / 0.20**	**0.02 / 0.03**
Copula: Clayton & Gamma	CRT [8]	0.00 / 0.00	0.00 / 0.00
	Distilled-CRT [19]	0.13 / 0.19	**0.07 / 0.07**
	Gaussian Mirror [45]	**0.02 / 0.18**	0.12 / 0.18
	Data Splitting [9]	**0.07 / 0.28**	0.04 / 0.02
	HRT [40]	0.00 / 0.00	0.00 / 0.00
	Deep Knockoff [30]	0.50 / 0.94	0.78 / 0.89
	DDLK [38]	0.00 / 0.00	0.00 / 0.00
	KnockoffGAN [14]	0.00 / 0.03	0.00 / 0.00
	sRMMD [22]	0.52 / 0.95	0.78 / 0.92
	DeepDRK [34]	0.01 / 0.08	0.12 / 0.13
	G^2M (ours)	**0.03 / 0.18**	**0.10 / 0.10**
Copula: Joe & Exp.	CRT [8]	0.00 / 0.00	0.00 / 0.00
	Distilled-CRT [19]	0.03 / 0.09	0.00 / 0.00
	Gaussian Mirror [45]	**0.05 / 0.28**	**0.01 / 0.02**
	Data Splitting [9]	0.06 / 0.19	**0.02 / 0.01**
	HRT [40]	0.00 / 0.00	0.00 / 0.00
	Deep Knockoff [30]	0.53 / 0.90	0.77 / 0.88
	DDLK [38]	0.00 / 0.00	0.00 / 0.00
	KnockoffGAN [14]	0.00 / 0.00	0.00 / 0.00
	sRMMD [22]	0.42 / 0.79	0.75 / 0.83
	DeepDRK [34]	0.09 / 0.21	0.08 / 0.00
	G^2M (ours)	**0.06 / 0.27**	0.02 / 0.01
Copula: Joe & Gamma	CRT [8]	0.00 / 0.00	0.00 / 0.00
	Distilled-CRT [19]	0.04 / 0.07	0.00 / 0.00
	Gaussian Mirror [45]	**0.09 / 0.52**	0.01 / 0.02
	Data Splitting [9]	0.07 / 0.28	0.02 / 0.01
	HRT [40]	0.00 / 0.00	0.00 / 0.00
	Deep Knockoff [30]	0.50 / 0.97	0.00 / 0.00
	DDLK [38]	0.01 / 0.00	0.00 / 0.00
	KnockoffGAN [14]	0.06 / 0.27	0.03 / 0.04
	sRMMD [22]	0.48 / 0.96	0.79 / 0.95
	DeepDRK [34]	0.05 / 0.30	**0.03 / 0.05**
	G^2M (ours)	**0.09 / 0.62**	**0.05 / 0.05**

Table 8: Extension of Table 1 with different noise levels: FDR / power under two additional different signal strength parameters $\frac{p}{10\sqrt{n}}$ and $\frac{p}{20\sqrt{n}}$. Red entries indicate FDR>0.1; among those with FDR ≤ 0.1, **Bold** entries indicate the highest power and **Blue** entries indicate the second highest.

Method	$\frac{p}{10\sqrt{n}}$		$\frac{p}{20\sqrt{n}}$	
	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.6$	$\rho = 0.7$
CRT [8]	0.04 / 0.96	**0.05 / 0.88**	0.07 / 0.42	0.09 / 0.25
Distilled-CRT [19]	0.40 / 1.00	0.40 / 1.00	0.26 / 0.89	0.21 / 0.65
Gaussian Mirror [45]	**0.09 / 1.00**	0.06 / 0.88	0.07 / 0.52	0.06 / 0.40
Data Splitting [9]	0.11 / 0.81	0.07 / 0.77	**0.03 / 0.56**	**0.08 / 0.45**
HRT [40]	0.01 / 0.71	0.02 / 0.46	0.02 / 0.13	0.00 / 0.05
Powerful Knockoff [37]	0.08 / 0.78	0.02 / 0.53	0.05 / 0.20	0.02 / 0.06
G^2M (ours)	**0.07 / 1.00**	**0.05 / 0.93**	**0.02 / 0.58**	**0.02 / 0.45**

where ϵ follows the standard normal distribution and the $m = 20$ for 20 covariates that are sampled uniformly.

G Supplementary Material for the Case Study

G.1 Nominal Metabolite List

In Table 11, we include the list of the nominal metabolites curated by [34] for the IBD dataset.

G.2 Additional Results for the IBD Study

Here we provide the supplementary information for the experimental results described in Section 4.3. In Table 12 and 13, we provide the list of identified metabolites by each of the considered models. This table provides additional information for Table 4 in the main paper which only includes metabolite counts due to limited space.

Table 9: Extension of Table 2 with different noise levels: FDR and power under two signal strength regimes $\frac{p}{10\sqrt{n}}$ and $\frac{p}{20\sqrt{n}}$. Red entries indicate FDR> 0.1; **Bold** for the highest power among methods controlling FDR ≤ 0.1; and **Blue** for the second highest.

Dataset	Method	$\frac{p}{10\sqrt{n}}$ (FDR / Power)	$\frac{p}{20\sqrt{n}}$ (FDR / Power)
Gaussian Mixture	CRT [8]	0.31 / 0.40	0.30 / 0.40
	Gaussian Mirror [45]	**0.06 / 1.00**	0.07 / 0.62
	Data Splitting [9]	**0.09 / 0.94**	**0.06 / 0.64**
	HRT [40]	0.01 / 0.39	0.02 / 0.09
	Deep Knockoff [30]	0.56 / 0.99	0.76 / 1.00
	DDLK [38]	0.79 / 1.00	0.72 / 0.91
	KnockoffGAN [14]	0.21 / 1.00	0.52 / 0.98
	sRMMD [22]	0.60 / 1.00	0.77 / 1.00
	DeepDRK [34]	**0.06 / 0.94**	0.15 / 0.70
	G^{2}M (ours)	**0.06 / 1.00**	**0.05 / 0.78**
Copula: Clayton & Exp.	CRT [8]	0.01 / 0.26	0.02 / 0.12
	Gaussian Mirror [45]	**0.05 / 0.95**	**0.08 / 0.46**
	Data Splitting [9]	0.08 / 0.86	0.13 / 0.46
	HRT [40]	0.02 / 0.58	0.00 / 0.06
	Deep Knockoff [30]	0.09 / 0.92	0.38 / 0.82
	DDLK [38]	0.49 / 0.88	0.02 / 0.05
	KnockoffGAN [14]	0.06 / 0.56	**0.05 / 0.12**
	sRMMD [22]	0.10 / 0.92	0.37 / 0.81
	DeepDRK [34]	0.07 / 0.85	0.17 / 0.51
	G^{2}M (ours)	**0.04 / 0.97**	0.12 / 0.59
Copula: Clayton & Gamma	CRT [8]	0.00 / 0.19	0.00 / 0.04
	Gaussian Mirror [45]	**0.09 / 1.00**	0.05 / 0.69
	Data Splitting [9]	0.10 / 0.96	**0.09 / 0.72**
	HRT [40]	0.02 / 0.92	0.02 / 0.29
	Deep Knockoff [30]	**0.07 / 0.97**	0.37 / 0.94
	DDLK [38]	0.47 / 0.95	0.14 / 0.30
	KnockoffGAN [14]	0.08 / 0.91	0.09 / 0.43
	sRMMD [22]	**0.06 / 0.97**	0.35 / 0.94
	DeepDRK [34]	0.06 / 0.92	0.12 / 0.67
	G^{2}M (ours)	**0.07 / 1.00**	**0.10 / 0.86**
Copula: Joe & Exponential	CRT [8]	0.11 / 0.57	**0.08 / 0.23**
	Gaussian Mirror [45]	0.05 / 0.76	0.15 / 0.35
	Data Splitting [9]	**0.09 / 0.78**	0.20 / 0.36
	HRT [40]	0.01 / 0.39	0.00 / 0.02
	Deep Knockoff [30]	0.18 / 0.91	0.48 / 0.79
	DDLK [38]	0.25 / 0.57	0.02 / 0.02
	KnockoffGAN [14]	0.05 / 0.37	**0.03 / 0.06**
	sRMMD [22]	0.16 / 0.88	0.40 / 0.68
	DeepDRK [34]	0.08 / 0.74	0.20 / 0.39
	G^{2}M (ours)	**0.08 / 0.91**	0.19 / 0.43
Copula: Joe & Gamma	CRT [8]	0.11 / 0.56	0.10 / 0.33
	Gaussian Mirror [45]	**0.05 / 0.98**	**0.08 / 0.56**
	Data Splitting [9]	0.07 / 0.92	0.10 / 0.58
	HRT [40]	0.01 / 0.91	0.01 / 0.16
	Deep Knockoff [30]	0.06 / 0.96	0.35 / 0.91
	DDLK [38]	0.48 / 0.95	0.09 / 0.17
	KnockoffGAN [14]	0.05 / 0.77	0.07 / 0.30
	sRMMD [22]	**0.06 / 0.95**	0.38 / 0.92
	DeepDRK [34]	0.05 / 0.86	0.12 / 0.61
	G^{2}M (ours)	**0.09 / 1.00**	**0.08 / 0.68**

Table 10: Feature selection performance on the Breast Cancer Wisconsin (Diagnostic) dataset [43]. Since the true support is unknown, we report "number of literature-referenced features / number of identified features" instead of FDR or power.

Model	G^2M (ours)†	CRT	Distilled-CRT	Gaussian Mirror†	Data Splitting†	HRT	Powerful Knockoff
Referenced / Identified	18/21	5/6	12/15	18/26	11/15	2/4	21/26
Model	DeepDRK	Deep Knockoff	sRMMD	KnockoffGAN	DDLK		
Referenced / Identified	15/20	11/15	13/18	17/22	8/13		

Table 11: IBD-associated metabolites that are supported by the literature. This table includes all 47 referenced metabolites for the IBD case study. Each metabolite is supported by one of the three sources: PubChem, peer-reviewed publications, or preprints. For PubChem case, we report the PubChem reference ID (CID), and for the other two cases, we report the publication references.

Reference Type	Metabolite	Source	Metabolite	Source
PubChem	palmitate	CID: 985	taurocholate	CID: 6675
	cholate	CID: 221493	p-hydroxyphenylacetate	CID: 127
	linoleate	CID: 5280450	deoxycholate	CID: 222528
	taurochenodeoxycholate	CID: 387316		
Publications	12.13-diHOME	[6]	dodecanedioate	[6]
	arachidonate	[6]	eicosatrienoate	[6, 5]
	eicosadienoate	[6]	docosapentaenoate	[6, 5]
	taurolithocholate	[6]	salicylate	[6]
	saccharin	[6]	1.2.3.4-tetrahydro-beta-carboline-1.3-dicarboxylate	[6]
	oleate	[5]	arachidate	[5]
	glycocholate	[5]	chenodeoxycholate	[5]
	phenyllactate	[22, 16]	glycolithocholate	[5]
	urobilin	[22, 28]	caproate	[22, 17]
	hydrocinnamate	[22, 15]	myristate	[22, 11]
	adrenate	[22, 20]	olmesartan	[22, 31]
	tetradecanedioate	[39, 23]	hexadecanedioate	[39, 23]
	oxypurinol	[7]	porphobilinogen	[24]
	caprate	[35, 36]	undecanedionate	[17, 42]
	stearate	[2, 5]	oleanate	[27]
	glycochenodeoxycholate	[33]	sebacate	[17]
	nervonic acid	[41]	lithocholate	[5]
Preprints	alpha-muricholate	[25]	tauro-alpha-muricholate/tauro-beta-muricholate	[25]
	17-methylstearate	[25]	myristoleate	[25]
	taurodeoxycholate	[25]	ketodeoxycholate	[25]

Table 12: A list of literature-supported metabolites out of a total of 80 candidates. "$*$" indicates the important metabolites marked by the corresponding algorithms.

Metabolite	$\mathbf{G^2M}$†	CRT	Distilled-CRT	Gaussian Mirror†	Data Splitting†	HRT	Powerful Knockoff
12.13-diHOME		*			*		
9.10-diHOME							
caproate		*		*	*	*	
hydrocinnamate				*	*		
mandelate		*					
3-hydroxyoctanoate				*			
caprate			*				
indoleacetate							
3-hydroxydecanoate							
dodecanoate			*				
undecanedionate		*					
myristoleate				*			
myristate							
dodecanedioate							
pentadecanoate		*		*			
hydroxymyristate							
palmitoleate							*
palmitate				*			
tetradecanedioate	*	*		*	*		
10-heptadecenoate							
2-hydroxyhexadecanoate	*						
alpha-linolenate			*				
linoleate	*		*	*			
oleate	*		*				
stearate	*	*				*	
hexadecanedioate				*	*		
10-nonadecenoate							
nonadecanoate	*	*	*			*	
17-methylstearate					*	*	
eicosapentaenoate							
arachidonate	*	*		*	*		*
eicosatrienoate			*				
eicosadienoate	*				*		*
eicosenoate	*		*				
arachidate			*				
phytanate							
docosahexaenoate			*		*		
docosapentaenoate	*	*			*		
adrenate	*		*	*	*		
13-docosenoate					*		
eicosanedioate		*	*		*	*	
oleanate			*				
masilinate				*			
lithocholate	*						
chenodeoxycholate	*		*				
deoxycholate							
hyodeoxycholate/ursodeoxycholate			*				
ketodeoxycholate			*				
alpha-muricholate							
cholate	*		*				
glycolithocholate							
glycochenodeoxycholate	*	*					
glycodeoxycholate	*						
glycoursodeoxycholate		*	*	*	*		
glycocholate	*			*			
taurolithocholate			*				
taurochenodeoxycholate	*	*		*	*		
taurodeoxycholate	*			*			
tauro-alpha-muricholate/tauro-beta-muricholate		*					
taurocholate	*		*	*			
salicylate		*	*	*	*		*
saccharin					*		
azelate							
sebacate							
carboxyibuprofen							
olmesartan							
1.2.3.4-tetrahydro-beta-carboline-1.3-dicarboxylate					*		
4-hydroxystyrene	*	*		*	*		*
acetytyrosine							
alpha-CEHC							
carnosol		*		*			
oxypurinol							
palmitoylethanolamide		*		*			
phenyllactate	*		*	*	*		*
p-hydroxyphenylacetate			*				
porphobilinogen				*	*		
urobilin		*	*	*	*		
nervonic acid			*	*	*		
oxymetazoline		*	*				

Table 13: A continued list of literature-supported metabolites out of a total of 80 candidates to Table 12. "$*$" indicates the important metabolites marked by the corresponding algorithms.

Metabolite	DeepDRK	Deep Knockoff	sRMMD	KnockoffGAN	DDLK
12.13-diHOME				$*$	
9.10-diHOME					
caproate	$*$	$*$	$*$		$*$
hydrocinnamate					
mandelate					
3-hydroxyoctanoate					
caprate					
indoleacetate					$*$
3-hydroxydecanoate					
dodecanoate				$*$	
undecanedionate	$*$			$*$	
myristoleate					
myristate					
dodecanedioate				$*$	
pentadecanoate					
hydroxymyristate					
palmitoleate					
palmitate				$*$	
tetradecanedioate		$*$			
10-heptadecenoate					
2-hydroxyhexadecanoate					
alpha-linolenate					$*$
linoleate					
oleate					
stearate					$*$
hexadecanedioate		$*$		$*$	$*$
10-nonadecenoate					
nonadecanoate					
17-methylstearate	$*$	$*$			$*$
eicosapentaenoate	$*$	$*$			$*$
arachidonate	$*$	$*$		$*$	$*$
eicosatrienoate	$*$	$*$			$*$
eicosadienoate	$*$	$*$	$*$		$*$
eicosenoate					
arachidate				$*$	
phytanate					
docosahexaenoate	$*$	$*$			$*$
docosapentaenoate	$*$	$*$		$*$	$*$
adrenate	$*$	$*$	$*$	$*$	$*$
13-docosenoate					
eicosanedioate	$*$	$*$			
oleanate					
masilinate					
lithocholate	$*$				
chenodeoxycholate					
deoxycholate	$*$			$*$	$*$
hyodeoxycholate/ursodeoxycholate					
ketodeoxycholate	$*$				
alpha-muricholate	$*$				
cholate		$*$			
glycolithocholate	$*$				
glycochenodeoxycholate					
glycodeoxycholate					
glycoursodeoxycholate					
glycocholate					
taurolithocholate					$*$
taurochenodeoxycholate					
taurodeoxycholate					
taurohyodeoxycholate/tauroursodeoxycholate					
tauro-alpha-muricholate/tauro-beta-muricholate		$*$			$*$
taurocholate					
salicylate	$*$	$*$	$*$	$*$	
saccharin				$*$	
azelate					$*$
sebacate	$*$				$*$
carboxyibuprofen					
olmesartan					
1.2.3.4-tetrahydro-beta-carboline-1.3-dicarboxylate					
4-hydroxystyrene		$*$		$*$	$*$
acetytyrosine					
alpha-CEHC					
carnosol					$*$
oxypurinol					
palmitoylethanolamide					
phenyllactate	$*$	$*$	$*$		$*$
p-hydroxyphenylacetate	$*$	$*$			$*$
porphobilinogen	$*$				
urobilin	$*$	$*$		$*$	$*$
nervonic acid					
oxymetazoline	$*$	$*$			$*$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly stated the objective on proposing a new method in FDR-controlled feature selection regime.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation is essentially described in the form of assumptions in the theoretical parts in the method section. We have also discussed the limitation in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Please refer to the method section for the theoretical results and the appendix for the proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper presents two algorithms that should provide full details of implementation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: `https://github.com/skyve2012/G2M`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Such details are introduced and discussed in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: As per feature selection, we follow the convention and report FDR and power (as empirical estimation of the means), with independent and repeated experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The configuration is specified in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The social impact and the importance of the work is discussed in the introduction section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: N/A.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing works/codes are all properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The assets are theoretical results and codes. The latter will be released upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No such assets involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No such assets involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is only used for grammatical correction and text polishing levels.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.