Journal of Machine Learning Research 24 (2023) 1-41 Submitted 10/22; Revised 3/23; Published 4/23

DART: Distance Assisted Recursive Testing

Xuechan Li XUECHAN.LI@DUKE.EDU
Department of Biostatistics
Duke University
Durham, NC 27705, USA

Anthony D. Sung ANTHONY.SUNG@DUKE.EDU
Department of Medicine
Duke University
Durham, NC 27705, USA

Jichun Xie JICHUN.XIE@DUKE.EDU
Department of Biostatistics
Duke University
Durham, NC 27705, USA

Editor: Sayan Mukherjee

Abstract

Multiple testing is a commonly used tool in modern data science. Sometimes, the hypotheses are embedded in a space; the distances between the hypotheses reflect their co-null/co-alternative patterns. Properly incorporating the distance information in testing will boost testing power. Hence, we developed a new multiple testing framework named Distance Assisted Recursive Testing (DART). DART features in joint artificial intelligence (AI) and statistics modeling. It has two stages. The first stage uses AI models to construct an aggregation tree that reflects the distance information. The second stage uses statistical models to embed the testing on the tree and control the false discovery rate. Theoretical analysis and numerical experiments demonstrated that DART generates valid, robust, and powerful results. We applied DART to a clinical trial in the allogeneic stem cell transplantation study to identify the gut microbiota whose abundance was impacted by post-transplant care.

Keywords: multiple testing, hierarchical testing, aggregation tree, false discovery rate, auxiliary information

1. Introduction

Multiple testing is commonly used to discover important features in modern data science. Each feature represents a hypothesis: the important features correspond to alternative hypotheses, and the rest to null hypotheses. A rejected hypothesis corresponds to an identified feature. The goal is to discover the alternative hypotheses with a controlled false discovery rate (FDR), the expected number of false discoveries over the total number of discoveries.

Under many circumstances, the hypotheses are coupled with other attributes in a metric space with a known pairwise distance. For example, previous brain studies have shown that the anatomical distance between the neurons can partially inform brain activities and neurons'

co-functioning patterns (Alexander-Bloch et al., 2013; Perinelli et al., 2019; Kristanto et al., 2020). For another example, the polygenic distance between amplicon sequence variants (ASVs) often informs their functional similarity and survival (Chen et al., 2012; Garcia et al., 2014; Martiny et al., 2015). In these examples, if we form neurons or ASVs as hypotheses, properly incorporating anatomy distance or the polygenic distance between hypotheses can improve the power in identifying functionally important neurons or ASVs. In other words, the distance among the hypotheses partially reflects the hypotheses' co-null or co-alternative status, called co-status hereafter.

We developed a new hierarchical multiple testing procedure called DART. It incorporates distance information to boost the power of testing. DART has two stages. The first stage is based on AI modeling: we use the automatic algorithm to construct an aggregation tree that incorporates the distance information and facilitates downstream testing. The second stage is based on statistical modeling: we perform a bottom-up multiple testing procedure on the aggregation tree to control FDR. Unlike traditional multiple testing that only uses statistical modeling, DART combines the power of statistical models and AI models to improve testing accuracy: statistical models are less flexible and less data-adaptive, but they can provide highly interpretable results; AI models fail to provide interpretable results, but they can explore complex underlying structures. Our study provides a new solution to generate data-adaptive and highly-interpretable results.

Our work is closely related to three streams of research.

- *Testing using distance/side information.* Some methods incorporate side information via parametric modeling. For example, Shu et al. (2015) imposes a 3D hidden Ising model to model the nearby hypothesis co-status. Lee and Lee (2016) uses a scalar parameter in a specific exponential-family distribution to control the level of co-status among nearby hypotheses. Cai et al. (2020) uses kernel functions to enforce similar prior null probabilities on nearby hypotheses. Lei and Fithian (2018) proposes an iterative FDR control procedure that incorporates side information via a parametric model estimated by the EM algorithm. These methods use parametric forms to model how distance/side information impacts the hypothesis co-status. Although parametric methods usually achieve good performances on simulated data, their performance on real data is unclear. For example, for fMRI analysis, Eklund et al. (2012) used extensive real data sets to show that many commonly used parametric models that specify temporal correlations in fMRI analysis are inappropriate, and thus lead to highly inflated type I error in multiple testing. Recently, Ramdas et al. (2019b) develops an FDR control procedure that allows scientists to incorporate four types of prior knowledge simultaneously. The procedure allows mix and match techniques and using multiple different forms of prior knowledge simultaneously while maintaining internal consistency among the pattern of rejections and acceptances. Besides, Xia et al. (2017) and Tansey et al. (2018) employ neural networks to leverage side information to improve testing accuracy while controlling FDR. All these methods specify the side information for each hypothesis specifically. However, the distance information is not defined for each hypothesis but for each hypothesis pair, and thus cannot be incorporated directly. Alternatively, some methods adopt non-parametric models to incorporate distance information into multiple testing. Zhang et al. (2011) uses a local neighborhood size and uses the aggregated

P-values in the neighborhood to perform multiple testing, called FDR_L. Li et al. (2013) uses the non-parametric propagation-separation approach (Belomestny and Spokoiny, 2007) to smooth the coefficients in the generating generalized estimating equations (GEE) models. These methods are more data adaptive. Because Li et al. (2013) only works for the GEE models, we will mainly compare DART with FDR_L later.

- *Hierarchical multiple testing.*

  a) *Gatekeeping.* Suppose the hypotheses are grouped into $m \geq 2$ ordered families. Gatekeeping procedures test a later family only if they reject the previous families (Dmitrienko et al., 2007, 2008, 2011; Dmitrienko and Tamhane, 2013; Xi and Tamhane, 2014) . The aim is to control the family-wise error rate (FWER). Gatekeeping procedures are usually designed for clinical trials. For other applications, the tests lose power when they discard the families of hypotheses whose previous family is accepted.

  b) *Top-down hierarchical testing.* Soriano and Ma (2017) applied a tree-structured Markov prior distribution to the indicators of hypotheses status and then calculated their posterior being alternative. The method relies on parametric modeling. Yekutieli (2008) proposes a top-down testing rule similar to gatekeeping. A node on the tree (a set with multiple hypotheses) will be tested only if its parent node is rejected. Other top-down multiple testing procedure also have been proposed for hypotheses structured in DAG, such as Ramdas et al. (2019a),Loper et al. (2022), Meijer and Goeman (2015), and Goeman and Finos (2012). These methods require the hypotheses to be partially ordered in the DAG, and thus not applicable to general hypotheses testing. Lei et al. (2020) proposes an iterative testing algorithm to perform FDR control on a series of contiguous candidate sets in a constrained set. However, for hypotheses with distance information, how to turn the distance into contiguous candidate sets is unclear.

  c) *Bottom-up hierarchical testing.* This approach tests the individual hypotheses first and then tests the aggregated hypotheses later. Our method, DART, adopts this approach. The most similar work to ours is the one introduced by Li et al. (2020b). They proposed a bottom-up procedure to test conjugate nodes (sets of hypotheses) with tree structures. A conjugate node is alternative only when all its containing hypotheses are alternative. Their method aims to control the node-level FDR. In contrast, DART focuses on each hypothesis. It aims to control hypothesis-level FDR. Thus, DART is fundamentally different.

- *Explainable AI (XAI).* XAI aims to (a) produce more explainable models while maintaining a high level of learning performance (prediction accuracy) and (b) enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI partners. (Turek, 2021) However, to the best of our knowledge, no XAI method has been proposed to control FDR in multiple testing. In general, pure AI modeling introduces intrinsic difficulties in controlling FDR.

2. Preliminaries

Suppose there are m hypotheses, forming the null and alternative hypothesis sets Ω_0 and Ω_1: $\Omega_0 \cap \Omega_1 = \emptyset$, $\Omega_0 \cup \Omega_1 = [m]$. For hypothesis i, a P-value statistic T_i is derived. Previous studies often assume the null P-values follow super-uniform distributions:

$$\forall\ i \in \Omega_0, \forall\ p \in (0,1), \mathrm{P}(T_i \le p) \le p.$$

Our work relaxes this assumption. We allow a null T_i asymptotically converges to a super-uniform statistic $\tilde{T}_i$ in the following way.

$$\sup_{i\in\Omega_0}\sup_{p\in\mathcal{P}_{i0}}\left|\frac{\mathrm{P}(T_i<p)}{\mathrm{P}(\tilde{T}_i<p)}-1\right| \le \delta_{0m}, \quad \text{where } \lim_{m\to\infty}\delta_{0m}=o(1)$$
$$\text{and } \mathcal{P}_{i0}=\left\{p\in[0,1]: P(\tilde{T}_i<p)\ge\left\{m(\log m\log\log m)^{1/2}\right\}^{-1}\right\}. \quad (1)$$

$\mathcal{P}_{i0}$ excludes the left tail regions (close to zero) to make the convergence easier. Otherwise, if $P(\tilde{T}_i < p)$ is too small, the convergence is hard to achieve. If $\tilde{T}_i$ follows a super-uniform (resp. uniform) distribution, we call T_i asymptotically super-uniform (resp. uniform). We relax the P-value null distribution assumptions because many P-values derived from asymptotic tests (*e.g.*, Wald, score, and likelihood-ratio tests) do not follow super-uniform distributions, but they are asymptotically super-uniform.

We do not make assumptions on alternative P-value distributions. Although they can be arbitrary, it is useful to think of the alternative P-values have larger probabilities to be small than the uniform distributions. Thus, many methods reject the hypotheses when the P-values are small. For example, a commonly used P-value threshold (Liu et al., 2013; Cai and Liu, 2016a; Xie and Li, 2018) is

$$\hat{t}=\sup\left\{\alpha_m\le t\le\alpha: \frac{mt}{\sum_{i\in[m]}I(T_i<t)}\le\alpha\right\}, \quad \text{where } \alpha_m=(m\log m)^{-1}. \quad (2)$$

Here, the dotted fraction notation denotes the shorthand $\frac{a}{b}=\frac{a}{b\vee 1}$. The threshold $\hat{t}$ for the Benjamini and Hochberg procedure (BH) (Benjamini and Hochberg, 1995) is slightly different but asymptotically equivalent to (2). In general, the rejection set is $\mathcal{R}=\{i: T_i\le t\}$ for some threshold t. Then, $\mathcal{U}=\Omega_1\cap\mathcal{R}$ is called the true discovery set, and $\mathcal{V}=\Omega_0\cap\mathcal{R}$ is called the false discovery set. The false discovery proportion and rate are defined as

$$\mathrm{FDP}:=\frac{|\mathcal{V}|}{|\mathcal{R}|}, \quad \mathrm{FDR}:=\mathrm{E}\left(\mathrm{FDP}\right),$$

where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$. Our task is to design a powerful rejection rule to asymptotically control FDR when m goes to infinity. We hope to gain power by properly incorporating the distance information between hypotheses, assuming they reflect their co-status patterns.

3. DART Algorithm Description

DART has two stages. Stage I uses an AI method to transfer the distance matrix into an aggregation tree, which defines the testing structure (Section 3.1). Stage II tests hypotheses embedded in the tree to gain power while controlling FDR (Section 3.2). An illustrating example is provided in Section 3.3.

3.1 Distance Matrix and Aggregation tree

Stage I aims to construct an aggregation tree $\mathcal{T}$ that provides a hierarchical testing structure for stage II. The aggregation tree $\mathcal{T}$ has L layers, constructed based on the distance matrix $\boldsymbol{D} = (d_{ij})$. The node-set on layer ℓ is denoted by $\mathcal{A}^{(\ell)}$. On the first (bottom) layer, $\mathcal{A}^{(1)} = \{\{1\}, \ldots, \{m\}\}$; each node represents a hypothesis. In general, $\mathcal{A}^{(\ell)}$ contains the nodes representing a set of hypotheses.

The hypotheses have different co-status patterns: some alternative hypotheses might stand-alone, some co-alternative within a small region, and others co-alternative within a large region. For any node A, denote its diameter (within-node distance) by $\mathrm{dia}(A) = \max_{i,j\in A} d_{ij}$. On higher layers of $\mathcal{T}$, nodes have larger diameters. We design multiple layers to construct nodes with various diameters and adapt to various co-status patterns.

In stage II, if a node is rejected, we will reject all the hypotheses within the node. For this purpose, we require the hypothesis distance within the node to be small so that they are likely co-status; thus, imposing the same decision rule on these hypotheses is reasonable. We require that for all $A \in \mathcal{A}^{(\ell)}$, $\mathrm{dia}(A) \le g^{(\ell)}$, where $g^{(\ell)}$ is a distance threshold with $0 < g^{(2)} < \ldots < g^{(L)}$. We set $g^{(1)} = 0$ because we do not need to aggregate hypotheses on the first layer.

Another requirement is to limit nodes' child numbers. Any node A in $\mathcal{A}^{(\ell)}$ with $\ell \ge 2$ is formed by aggregating the nodes from the previous layer. These nodes are called the children of A; they form the child set $\mathcal{C}(A)$. We require that $|\mathcal{C}(A)| \le M$ for any node A. Here, M is called the maximum node size. We set up this requirement to increase the following testing's stability: if A contains too many hypotheses, rejecting A leads to rejecting all hypotheses in A. This introduces large variability in testing and makes the algorithm outputs unstable.

To make sure all the nodes on layer ℓ satisfy $\mathrm{dia}(A) \le g^{(\ell)}$ and $|\mathcal{C}(A)| \le M$, we developed Algorithm 1. The key strategy is to recursively set $\mathcal{A}^{(\ell)}$ based on $\mathcal{A}^{(\ell-1)}$ and find the closest node pair for aggregation. The resulting tree depends on the tuning parameter L, M, and $g^{(\ell)}$ with $\ell \in \{2, \ldots, L\}$. We introduce the tuning parameter selection criteria in Section 3.4.

Algorithm 1 is a feasible algorithm to construct an aggregation tree satisfying $\mathrm{dia}(A) \le g^{(\ell)}$ and $|\mathcal{C}(A)| \le M$. Alternative AI approaches could also be used. For example, an aggregation tree can be constructed by recursively applying community detection algorithms, such as K-means clustering (Jin and Han, 2010), Louvain method (Blondel et al., 2008), and Leiden algorithm (Traag et al., 2019), in the similar spirit of hierarchical clustering (Zepeda-Mendoza and Resendis-Antonio, 2013). If a modified algorithm based on these algorithms could restrict the maximum node size, it may also be applied here. The key of the desired algorithm is to generate a tree with few mixed nodes (defined in Section 4) to ensure the asymptotic FDR control of Step 2.

3.2 Recursive Testing Embedded in the Tree

Recall that the tree nodes consist of close hypotheses likely to be co-status. We developed a multiple testing algorithm that incorporates the tree to improve the testing power.

Algorithm 2 describes the recursive testing procedure from the single-hypothesis layer (bottom) to the large-node layer (top). It starts with testing all hypotheses using the traditional FDR control procedure. This step does not use any hypothesis co-status patterns. Thus, DART is likely to discover the hypotheses with strong signal-to-noise ratios (SNRs) on

the bottom layer. On higher layers, DART tests larger nodes containing more hypotheses. The weaker-SNR hypotheses are likely to be aggregated then to increase their identification chances.

Algorithm 2 mentioned a few new terms: dynamic nodes, candidate node P-values, and recursive P-value cutoffs. We provide more details on these terms below.

Candidate dynamic nodes. The dynamic nodes are the nodes excluding the rejected hypotheses on previous layers. If a hypothesis is rejected, we will not test it again. There are two main reasons. First, if the rejected hypotheses were not removed, they could become non-significant after being aggregated with other null hypotheses on higher layers. It introduces interpretation difficulties. Second, a rejected hypothesis on low layers usually has strong SNRs. If we include these hypotheses in higher layers, this hypothesis alone may make the entire node significant, even if the node contains null hypotheses. Then, the rejection of the entire node may lead to false rejections on its containing null hypotheses.

A candidate dynamic node is a dynamic node with $|\mathcal{C}(S)| \geq 2$. If a node S has $|\mathcal{C}(S)| = 1$, this node must have been tested on previous layers. Thus, we do not need to test it again. The set of the candidate dynamic nodes on layer ℓ is denoted by $\mathcal{B}^{(\ell)}$.

Candidate node P-values. For any candidate dynamic node S, we use the Gaussian aggregation approach to derive candidate node P-value: $T_S = \bar{\Phi}\left\{\sum_{j\in S} \bar{\Phi}^{-1}(T_j)/\sqrt{|S|}\right\}$, where $\bar{\Phi}$ is the complement cumulative density function (CCDF) of the standard Gaussian distribution. The aggregation is efficient. Alternatively, we may consider using other aggregation approaches, such as the Fisher's combination (Fisher, 1925), the chi-square aggregation, and the Cauchy aggregation (Liu and Xie, 2020). The Fisher's combination and the chi-square aggregation approaches have lower power than the Gaussian aggregation when the hypotheses in the node are co-status. The Cauchy aggregation relies on the heavy-tail Cauchy density; thus, it introduces theoretical challenges to studying the asymptotic null distributions of the node P-values. The main challenge lies in accounting for the post-selection effect: the hypotheses and their P-values only have the chance to be aggregated when they are not rejected on the previous layers. In contrast, for Gaussian aggregation, the post-selection problem can be solved because the candidate node P-values are still asymptotically super-uniform (Lemmas 7 and 8). Thus, we go with the Gaussian aggregation.

P-value cutoffs. On layer ℓ, we already have the set of the rejected hypotheses on the previous $\ell - 1$ layers, denoted by $\mathcal{R}^{(1:\ell-1)}$. We set up the P-value threshold $\hat{t}^{(\ell)}$ as

$$\hat{t}^{(\ell)} = \sup\left\{\alpha_m \leq t \leq \alpha : \frac{\sum_{\ell'=1}^{\ell-1} m^{(\ell')}\hat{t}^{(\ell')} + m^{(\ell)}t}{|\mathcal{R}^{(1:\ell-1)}| + \sum_{S\in\mathcal{B}^{(\ell)}} |S| I(T_S < t)} \leq \alpha\right\}. \tag{3}$$

Here, $\alpha_m = (m \log m)^{-1}$ and $m^{(\ell')} = |\mathcal{B}^{(\ell')}|$. It is easy to see that $\hat{t}^{(\ell)}$ depends on the previous layer's cutoff and rejection set. Comparing $\hat{t}$ in (2) and $\hat{t}^{(\ell)}$ in (3), it is easy to see that they share some similarities. The numerators bound the false discoveries asymptotically, and the denominators count the total discoveries. By making the fraction less than or equal to the desired FDR level α, DART asymptotically controls the weighted node-FDR (Section 4). DART also asymptotically controls the hypothesis level FDR when most nodes contain co-status hypotheses.

Algorithm 1: Transform the distance matrix into an aggregation tree.

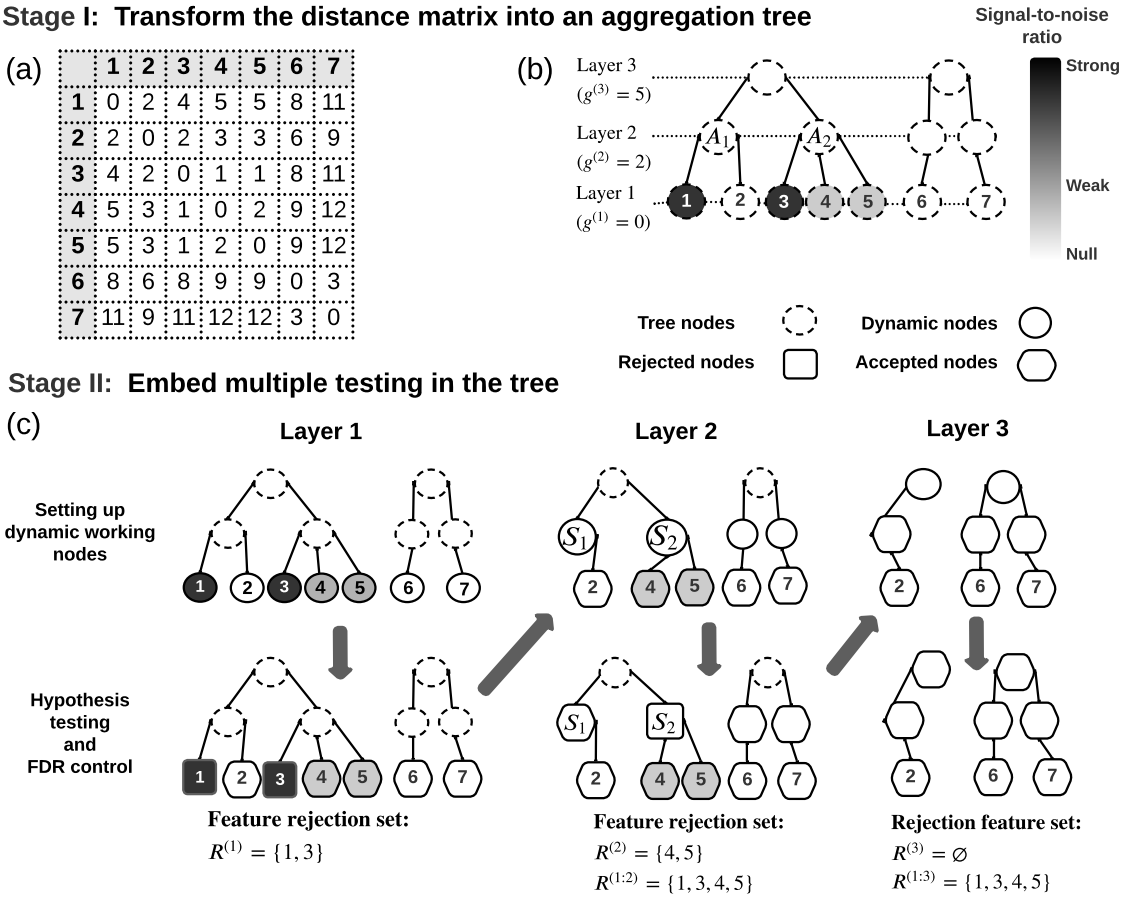
Data: Distance matrix $\boldsymbol{D} = (d_{ij})_{m\times m}$, maximum layer L, maximum node size M, distance thresholds $g^{(\ell)}$ with $\ell \in [L]$.
Result: An aggregation tree $\mathcal{T} = \{\mathcal{A}^{(\ell)} : \ell \in [L]\}$.
$\mathcal{A}^{(1)} = \{\{1\}, \ldots, \{m\}\}$;
for $\ell \in \{2, \ldots, L\}$ **do**
  $A^{(\ell)} = \mathcal{A}^{(\ell-1)}$; flag=TRUE ; *// Initiate $\mathcal{A}^{(\ell)}$*
  Calculate the between-node distances: $\forall\ A_1, A_2 \in \mathcal{A}^{(\ell)}$, $\mathrm{dist}(A_1, A_2) = \max_{i\in A_1, j\in A_2} d_{ij}$;
  while *flag=TRUE* **do**
    Pick up the node pair $(\breve{A}_1, \breve{A}_2)$ with the smallest between-node distance;
    if $\mathrm{dist}(\breve{A}_1, \breve{A}_2) > g^{(\ell)}$ **then** flag=FALSE ;
    else
      if $|\mathcal{C}(\breve{A}_1 \cup \breve{A}_2)| \le M$ **then**
        Put $\breve{A} = \breve{A}_1 \cup \breve{A}_2$ in $\mathcal{A}^{(\ell)}$ and remove $\breve{A}_1$ and $\breve{A}_2$;
        Set the between-node distance: $\forall\ A \in \mathcal{A}^{(\ell)}$, $\mathrm{dist}(A, \breve{A}) = \max_{i\in A, j\in \breve{A}} d_{ij}$;
      else Update the between-node distance: $\mathrm{dist}(\breve{A}_1, \breve{A}_2) = \infty$;

Algorithm 2: Recursive testing embedded in the tree.

Data: P-values $T_1, \ldots, T_m$; tree $\mathcal{T} = \{\mathcal{A}^{(\ell)} : \ell \in [L]\}$.
Result: Rejection set $\mathcal{R}$.
Following the BH procedure (2) to set the threshold $\hat{t}^{(1)}$ and $\mathcal{R} = I(i : T_i < \hat{t}^{(1)})$;
for $\ell \in \{2, \ldots, L\}$ **do**
  Define the **candidate dynamic node** set $\mathcal{B}^{(\ell)} = \{S : S = A \setminus \mathcal{R}, |\mathcal{C}(S) \ge 2|\}$;
  For all $S \in \mathcal{B}^{(\ell)}$, get the **candidate node P-values** T_S ;
  Set the recursive **P-value cutoff** $\hat{t}^{(\ell)}$ as in (3) and let $\mathcal{R}^{(\ell)}_{\text{node}} = \{S : T_S < \hat{t}^{(\ell)}\}$;
  Map the rejections to hypothses and update $\mathcal{R} = \mathcal{R} \cup \{i : i \in S, S \in \mathcal{R}^{(\ell)}_{\text{node}}\}$.

3.3 A Toy Example to Illustrate DART

We provide a toy example in Figure 1 to illustrate DART. The detailed algorithm descriptions are provided in Algorithm 1 and Algorithm 2 in Section 3.



	1	2	3	4	5	6	7
1	0	2	4	5	5	8	11
2	2	0	2	3	3	6	9
3	4	2	0	1	1	8	11
4	5	3	1	0	2	9	12
5	5	3	1	2	0	9	12
6	8	6	8	9	9	0	3
7	11	9	11	12	12	3	0

Figure 1: An illustrating example of DART with 7 features.

Figure 1(a) shows the distance matrix of the seven hypotheses. In stage I, we transfer the distance matrix into the 3-layer aggregation tree based on Algorithm 1. The resulting 3-layer aggregation tree is shown in Figure 1(b). Each leaf (node on the bottom layer) on the tree corresponds to a hypothesis and a test statistic. The gray scales illustrate the underlying SNR ratios in the statistics; these SNR ratios are unknown.

In stage II, we perform the multiple testing embedded in the aggregation tree based on Algorithm 2. We start from the bottom layer and hierarchically proceed to higher layers.

- On layer 1 (the bottom layer), hypotheses 1 and 3 are rejected because their test statistics have strong SNR ratios.

- After the testing procedure on each layer, the rejected nodes are marked by solid squares and the accepted nodes solid hexagons. If a node is rejected, all its containing hypotheses are rejected.

- When testing on a higher layer, all previously rejected features are excluded from the nodes (dashed-line circled) to form the dynamic nodes (solid-line circled) on this layer. For example, on layer 2, tree node $A_1 = \{1, 2\}$ turns into the dynamic node $S_1 = \{2\}$ because hypothesis 1 is rejected on the bottom layer. S_1 will not be tested on layer two because it only contains hypothesis 2, already tested on layer 1. Using the terminology in Section 3.2, the number of children $|\mathcal{C}(S_2)| = 1$; thus, S_1 is not a candidate dynamic node and will not be tested on layer 2. For another example, $A_2 = \{3, 4, 5\}$ turns to the dynamic node $S_2 = \{4, 5\}$ on layer 2 because hypothesis 3 was rejected on the bottom layer. $|\mathcal{C}(S_2)| = 2$ and thus, it will be tested on layer 2.

- The test statistics of hypotheses 4 and 5 have relatively weak SNRs. They are not significant enough to be rejected on layer 1. However, on layer 2, they are aggregated to form the dynamic node $S_2 = \{4, 5\}$; the aggregated SNR is large enough so that S_2 is rejected, leading to the rejections of hypotheses 4 and 5. Thus, the power of DART is higher than the power of a single-layer testing method.

3.4 Tuning Parameter Selection

Proper parameters will result in a tree empowering testing. We suggest setting the maximum node size M as 2 or 3. Appendix C verified that when $M = 2$ or $M = 3$, DART asymptotically controls FDR and is more powerful. Denote the desired minimal number of nodes on layer L by c_m. If $c_m < 35$, DART's asymptotic validity might fail to kick in, leading to possibly inflated FDR. Therefore, we request $c_m \geq 35$. Accordingly, we set the maximum layer number $L = \lfloor \log_M m - \log_M c_m \rfloor \leq \lfloor \log_M m - \log_M 35 \rfloor$. The distance thresholds are set recursively to maximize the number of candidate nodes on each layer. We first try a set of possible thresholds $G = \{g_1, \ldots, g_K\}$. On layer ℓ with $\ell \geq 2$, we let $G^{(\ell)} = \{g \in G : g \geq g^{(\ell-1)}\}$. Hierarchically, on layer ℓ, we try every $g \in G^{(\ell)}$ and count the number of resulting candidate nodes on this layer. We set $g^{(\ell)}$ as the g with the most candidate nodes. See Algorithm 3 in Appendix C for details.

4. Asymptotic Validity

This section shows that DART asymptotically controls the hypothesis FDR under mild conditions.

Weighted node-FDR and hypothesis FDR. For any candidate dynamic node, if the node contains any alternative hypothesis, we call the node alternative; otherwise, it is called null. On layer ℓ, we denote the set of null candidate dynamic nodes by $\mathcal{B}_0^{(\ell)}$ and the set of rejected nodes by $\mathcal{R}_{\text{node}}^{(\ell)}$. Then the weighted node-FDR is

$$\text{FDP}_{\text{node}} := \frac{\sum_{\ell=1}^{L} \sum_{S \in \mathcal{R}_{\text{node}}^{(\ell)} \cap \mathcal{B}_0^{(\ell)}} |S|}{\sum_{\ell=1}^{L} \sum_{S \in \mathcal{R}_{\text{node}}^{(\ell)}} |S|} = \frac{\sum_{\ell=1}^{L} \sum_{S \in \mathcal{R}_{\text{node}}^{(\ell)} \cap \mathcal{B}_0^{(\ell)}} |S|}{|\mathcal{R}|}, \quad \text{FDR}_{\text{node}} = \text{E}(\text{FDP}_{\text{node}}).$$

In contrast, the hypothesis FDR is

$$\text{FDP} := \frac{|\mathcal{R}| \cap \Omega_0}{|\mathcal{R}|}, \quad \text{FDR} = \text{E}(\text{FDP}).$$

Notably, FDP_{node} only accounts for the false discoveries in null nodes. If an alternative node containing both null and alternative hypotheses is rejected, the rejection will not increase the numerator of FDP_{node} but will increase the numerator of FDP. Thus, $\text{FDP}_{\text{node}} \leq \text{FDP}$. Although our ultimate goal is to control FDP, we control FDP_{node} as an intermediate step. The difference between FDP_{node} and FDP relies on the number of the rejected mixed nodes that contain both null and alternative hypotheses.

Weighted node-FDR control. We require the following conditions for weighted node-FDR control.

- *Condition 1. Sparse alternatives.* The alternative hypothesis number $m_1 = O(m^{r_1})$, for some $r_1 < (M^{L-1}+1)^{-1}$.
- *Condition 2. Sufficient moderate SNR nodes (see Definition 4 in Appendix A).* Denote the number of moderate SNR nodes on the tree $\mathcal{T}$ by m_{md}. We require that $m_{\text{md}} \geq O(\log m)$. A moderate SNR node (a) contains no hypotheses that will be rejected with non-vanishing probabilities before its locating layer, and (b) will be rejected on its locating layer with a non-vanishing positive probability. The existence of these nodes is to guarantee that some alternative nodes are rejected on each layer so that the threshold $\hat{t}^{(\ell)}$ is not too small; otherwise, the number of total rejections will be too small so that a single false rejection would inflate FDR.
- *Condition 3. Almost independence.* Most hypothesis P-values are mutually independent. The number of dependent P-values does not exceed $o(m_{\text{md}})$.

Here, Conditions 1 and 2 are inherited and extended from the previous multiple testing literature (Cai and Liu, 2016b). Condition 1 assumes the alternative hypothesis sparsity. Condition 2 usually holds when the sample size n is sufficiently large compared to p, L is properly chosen, and the signal-to-noise ratio distribution of the alternative hypotheses has continuous support over a large range. Condition 3 is a strong assumption. We require it to ensure that after higher-layer aggregation, most node P-values are still asymptotically super-uniformly distributed under the null. It is possible to relax this condition. However, the proof will be much more complicated.

Lemma 1 *Under Conditions 1–3, at any pre-specified level $\alpha \in (0,1)$, Algorithm 2 satisfies the following asymptotic validity.*

(a) For any $\epsilon > 0$, $\lim_{m\to\infty} \text{P}(FDP_{\text{node}} \leq \alpha + \epsilon) = 1$. Consequently, $\lim_{m\to\infty} FDR_{\text{node}} \leq \alpha$.

(b) Let $\tilde{\Omega}_0$ be the set of null P-values that are asymptotically uniform. If $\lim_{m\to\infty} |\tilde{\Omega}_0|/m = 1$, then for any $\epsilon > 0$, $\lim_{m\to\infty} \text{P}(|FDP_{\text{node}} - \alpha| \leq \epsilon) = 1$. Consequently, $\lim_{m\to\infty} FDR_{\text{node}} = \alpha$.

See Appendix B for proof of this lemma. Two main challenges in the proof are the hierarchical testing structure and the post-selective effect introduced by the dynamic nodes. Thus, we proved the FDR_{node} control recursively, starting from the bottom layer. The bottom layer follows the BH procedure. Then, given the FDR_{node} control on the previous layers, we proved the control on the current layer. Recall that dynamic nodes do not contain the already-rejected hypotheses. To account for the post-selection effect, we proved that

conditioning on the testing results from the previous layers, the dynamic node P-values are still asymptotically super-uniform or asymptotically uniform.

Hypothesis FDR control. Previously, we constructed a tree where the hypotheses were hierarchically aggregated into the nodes based on their distance. Thus, we expect many nodes contain co-status hypotheses. However, some large nodes on high layers may be mixed, containing null and alternative hypotheses. Some mixed nodes are concerning, while others are not. For example, suppose a node A on layer ℓ contains null hypotheses, strong SNR hypotheses (see Definition 5 in Appendix A) and weak SNR hypotheses (see Definition 6 in Appendix A). The strong SNR hypotheses are probably rejected before layer ℓ. Thus, when Algorithm 2 reaches layer ℓ, A probably already turns into a dynamic node only containing the null and weak SNR hypotheses. The weak SNR hypotheses are the alternative hypotheses unlikely to be rejected. As a result, the null hypotheses in A will not be rejected, and thus, the existence of A will not inflate the FDR. Thus, to control FDR, we only need to restrict those concerning mixed nodes whose null hypotheses are likely to be rejected with non-vanishing probabilities.

Definition 2 (Concerning mixed nodes) *For any node $A \in \mathcal{A}^{(L)}$, let*

$$A^* = A \setminus (\Omega_{st} \cup \Omega_{wk}),$$

where Ω_{st} is the strong SNR hypothesis set, and Ω_{wk} is the weak SNR hypothesis set. The definitions of Ω_{st} and Ω_{wk} are provided in Appendix A. If $A^* \cap \Omega_0 \neq \emptyset$ and $A^* \cap \Omega_1 \neq \emptyset$, we call A a concerning mixed node.

- *Condition 4. sparse concerning mixed nodes.* On the top layer of the tree $\mathcal{A}^{(L)}$, the number of the concerning mixed nodes cannot exceed $o(m_{\text{md}})$.

We allow the existence of concerning mixed nodes, but to asymptotically control FDR, Algorithm 2 cannot afford too many of them. Condition 4 specifies the tolerance level. Intrinsically, Condition 4 depends on the assumption that the distance matrix predominantly reflects the hypothesis co-status. If so, with properly selected $\{g^{(\ell)} : \ell \in [L]\}$, Algorithm 1 will probably generate a tree satisfying Condition 4, because it uses the greedy algorithm to aggregate the closest remaining hypotheses. On the other hand, if this assumption does not hold, Algorithm 1 cannot generate a tree with nodes implying hypothesis co-status. Under this case, we do not recommend using DART.

By adding Condition 4, we extend the FDR_{node} control to FDR control (Theorem 3).

Theorem 3 *Under Conditions 1-4, at any pre-specified level $\alpha \in (0, 1)$, Algorithm 2 satisfies the following asymptotic vality.*

(a) For any $\epsilon > 0$, $\lim_{m\to\infty} \mathrm{P}(FDP \leq \alpha + \epsilon) = 1$. Consequently, $\lim_{m\to\infty} FDR \leq \alpha$.

(b) Let $\tilde{\Omega}_0$ be the set of null P-values that are asymptotically uniform. If $\lim_{m\to\infty} |\tilde{\Omega}_0|/m = 1$, then for any $\epsilon > 0$, $\lim_{m\to\infty} \mathrm{P}(|FDP - \alpha| \leq \epsilon) = 1$. Consequently, $\lim_{m\to\infty} FDR = \alpha$.

5. Numerical Simulation

We simulated $m = 1000$ features located in the two-dimensional Euclidean space with randomly generated location coordinates: the first coordinate follows $\mathrm{N}(0, 2)$, and the second

coordinate follows Unif$(0, 4)$. A distance matrix $\boldsymbol{D} = (d_{ij})_{m\times m}$ was calculated based on the Euclidean distance between two features' locations. Feature i links to a parameter of interest θ_i. The hypotheses are $H_{0,i} : \theta_i = 0$ versus $H_{1,i} : \theta_i \neq 0$, $i \in [m]$.

We considered four settings, SE1–SE4. SE1 simulated a straightforward case where the P-values follow uniform distributions under the null. SE2 misspecified the null distributions of the test statistics, in order to evaluate the methods' robustness. SE3 simulated the linear regression model, and SE4 simulated the Cox proportional hazard model; their P-values were derived from the Wald tests. Each setting contained 200 repetitions. The setting details were described in Appendix C.

Under different nominal FDR levels $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$, we compared the performance of DART and its competitors: BH (Benjamini and Hochberg, 1995), AdaPT (Lei and Fithian, 2018) and FDR$_L$ (FDR$_L$ I and FDR$_L$ II) (Zhang et al., 2011). AdaPT incorporates the location coordinates as side information; DART incorporates the distance matrix; FDR$_L$ incorporates the information of each hypothesis's k nearest neighbors. Thus, the settings favor AdaPT because we provided it with the most information. The tuning parameters used in DART, AdaPT, and FDR$_L$ procedures were discussed in Appendix C.

Figure 2A shows the type I error (measured by average FDP) and power (measured by average sensitivity) and their error bars of various methods:

- *Average FDP*: Under SE1, SE3, and SE4, DART, BH, AdaPT, and FDR$_L$ II control the average FDP well. Under SE2, DART's (resp. BH's) average FDP is slightly inflated when $\alpha = 5\%$ (resp. $\alpha \leq 15\%$). This is because we deliberately misspecified the P-value null distributions in SE2. AdaPT has consistently good FDR control. In contrast, FDR$_L$ I exhibited severe FDR inflation under all four settings. FDR$_L$ I has longer error bars than DART; so does FDR$_L$ II when $\alpha \geq 10\%$. This suggests that DART's FDP is less variable than FDR$_L$ I and II.

- *Average sensitivity*: DART's sensitivities are consistently higher than BH. DART has much higher sensitivities than AdaPT (resp. FDR$_L$ II) when $\alpha \leq 15\%$ (resp. $\alpha \leq 10\%$) and slightly lower sensitivities when $\alpha = 20\%$. If a low nominal FDR level (such as 5%) is preferred, DART is the most powerful among all methods.

- *Computation time*: DART is computationally efficient. For example, on average, one run (per repetition) of DART takes only 0.64 minutes across all settings. In contrast, AdaPT failed in generating any testing results within 8 minutes in about 17% of the runs (Table 1 in Appendix C). Among AdaPT's successful runs (within 8 minutes), one run on average takes 3.90 minutes. DART is at least 6 times faster.

DART assumes that the distances reflect the hypothesis co-status patterns. However, in practice, this assumption could be partially violated. To assess the methods' robustness, we switched the proportion τ of the alternative hypotheses with the null hypotheses (Appendix C). Figure 2B shows that FDR$_L$ have inflated average FDP when the switching proportion $\tau \geq 6\%$. All other methods still have good FDR control. Even under these assumption partial violation cases, DART's sensitivity is still much higher than BH. Compared to AdaPT, DART still has higher sensitivity when $\alpha \leq 15\%$ and a slightly lower average sensitivity when $\alpha = 20\%$. These results show that DART's performance is consistently satisfying even when the data are less ideal.

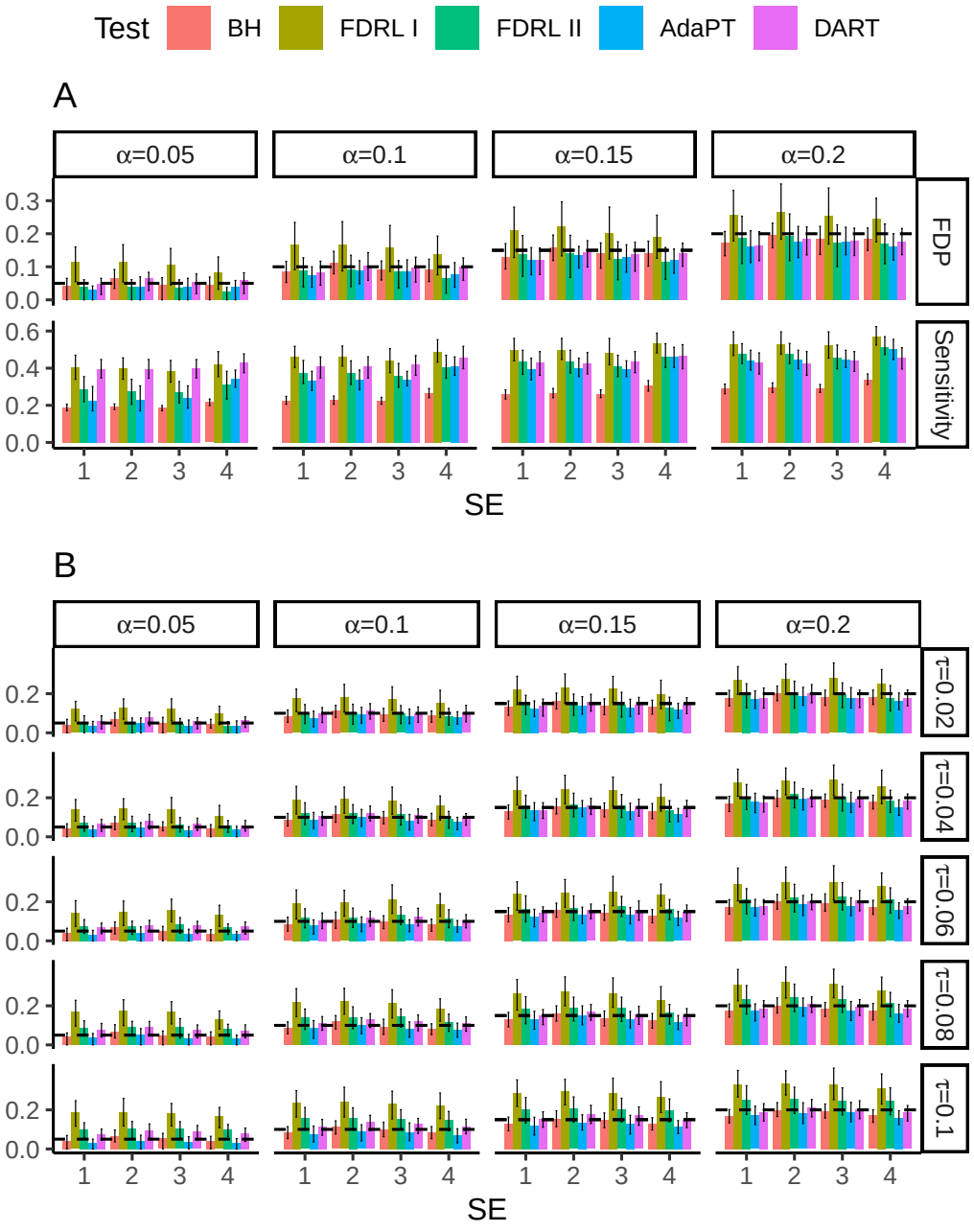


Figure 2: A, FDP and sensitivity of various testing methods under SE1-SE4. B, FDP of various testing methods under SE1-SE4 when proportion τ of the alternative hypotheses were switched with the null. The dashed lines in the FDP panels mark the nominal FDR levels.

One reason for DART's satisfying performance is because DART incorporates AI modeling to transfer the distance matrix into an aggregation tree, which later defines the testing structure. The AI modeling is data adaptive (compared to fixed neighborhood modeling as in FDR_L) and robust (compared to the parametric modeling as in AdaPT) and thus generate satisfying results under various settings.

6. Real-world Experiment

We applied DART to a clinical trial on hematopoietic stem cell transplantation (HCT). Graft-versus-host disease (GVHD) is one of the major complications of HCT. Recent studies have linked GVHD to the disruptions of the gut microbiome (Jenq et al., 2012). The disruptions may be related to the environmental changes such as post-transplant care (Claesson et al., 2012). This study investigates the impact of two post-transplant cares, home care (HC) and standard hospital care (SH), on patients' gut microbiota compositions.

In our data, patient fecal samples were collected before and after HCT; the fecal microbiome are sequenced by the 16S ribosomal RNA sequencing at the Memorial Sloan Kettering Cancer Center. The data were then pre-processed by the R package, DADA2 (Callahan et al., 2016), to generate the amplicon sequence variants (ASV) and the read counts. To improve the analysis quality, we removed the ASVs present in fewer than 10% of the samples. Samples with follow-up time longer than 1-year was also removed from the study. The zero counts were replaced by 0.5 (Aitchison, 1982; Kurtz et al., 2015).

After pre-processing, our data contain 456 microbiome samples from 126 leukemia patients before and after the HCT. Each microbiome sample contains 866 amplicon sequence variants (ASVs). We excluded the 9 ASVs with missing taxonomy order information. In microbiome studies, the ASV relative abundance (measured by its abundance proportions) is more meaningful than its absolute abundance. Thus, for the remaining 857 ASVs, we calculated their log odds. Here, the odds for an ASV is

$$\text{odds} = \frac{\text{ASV abundance proportion}}{1 - \text{ASV abundance proportion}}.$$

These ASV's abundance proportions do not add up to 1 because 9 ASVs were excluded. We set up the longitudinal linear mixed model

$$\text{Y}_{ijk} = \theta_{0,i} + \theta_{1,i} W_{1,k} + \theta_{2,i} W_{2,jk} + \theta_{3,i} W_{1,jk} W_{2,jk} + b_{ij} + \epsilon_{ijk}. \tag{4}$$

Here i is the ASV index, j is the sample index, and k is the patient index. The outcome Y_{ijk} is the log odds of ASV_i when sample j of patient k was collected. For patient k, $W_{1,k}$ is one's after-transplant care type (HC for 1 and SH for 0), $W_{2,jk}$ is care time length. b_{ij} is the random effect to incorporate the dependence across measurements for the same patient, and ϵ_{ijk} is the random error. To identify ASVs whose abundance change is impacted by the after-transplant care (the interaction between the post-transplant care type and time), we set up the hypotheses: $H_{0,i} : \theta_{3,i} = 0, i \in [857]$. The distances between the hypotheses were defined by the evolutionary distances among ASVs. Previous studies showed that evolutionally close ASVs might be functional similar. (Chen et al., 2012; Garcia et al., 2014; Martiny et al., 2015). We used the Wald tests to calculate the P-values.

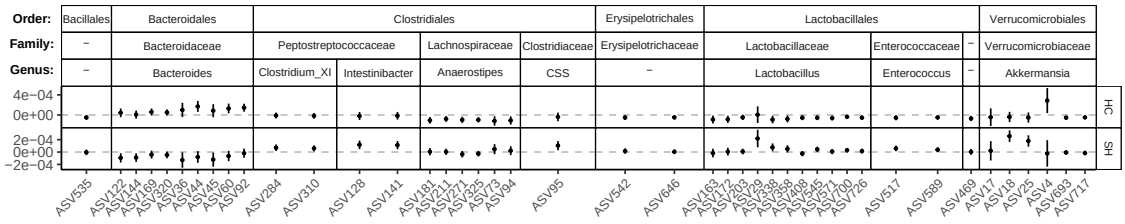


Figure 3: Forest plots to visualize the after-transplant time effect in home care group (HC) and standard hospital care group (SH) on the abundance of 43 DART-identified ASVs. A dot represents the estimated time effect on an ASV; a vertical line marks its 95% confidence interval. Here, CSS stands for Genus Clostridium sensu stricto.

We applied BH and DART to test the hypotheses with the nominal FDR level 5%. Details about aggregation tree construction in DART can be found in Appendix C. BH failed to identify any important ASVs. In contrast, DART identified 43 ASVs by incorporating the evolutionary distance information among ASVs. Among them, 39 ASVs have well-annotated Genus information. Figure 3 shows their log relative abundance change across time in home-care (HC) and standard hospital care (SH) groups Higher abundance in *Enterococcus*, *Clostridium XI* and *Akkermansia* were associated with more severe GVHD (Payen et al., 2020; Li et al., 2020a,a; Shono et al., 2016). Most of the ASVs in these three genera had less abundance over time in the home care group, suggesting home care might help reduce the GVHD severity. Higher abundances in *Bacteroides*, *Anaerostipes* and *Lactobacillus* were associated with reduced GVHD severity (Payen et al., 2020; Lin et al., 2021). Almost all the identified ASVs in *Bacteroides* had increased relative abundance in the home care group but decreasing abundance in the standard hospital care group. These findings suggested that home care might help relieve the GVHD severity.

7. Conclusion and Discussion

In this paper, we developed a novel multiple testing method, DART, to incorporate feature distance in multiple testing. Under many application contexts, the feature distances serve as auxiliary information of their co-importance pattern. DART incorporates this information to boost the testing power. DART applies to the P-values obtained from many asymptotic tests, and thus can work with a wide range of models.

Stage 1 of DART involves constructing an aggregation tree. We provided Algorithm 1 to construct the aggregation tree. Other algorithms may also work, and result in a different aggregation tree from the same distance matrix. Consequently, Stage 2 testing process could lead to different results based on different trees. In practice, if several aggregation trees exist, DART can be applied to all of them, and we can take the one with the most rejections. The asymptotic validity will still hold for this procedure.

The main limitation of the work is that the FDR control is asymptotic and relies on several conditions. Recently, many hypothesis testing literature develops finite-sample FDR control procedures. These procedures usually impose stronger assumptions on p-values

or test-statistics (Lei and Fithian, 2018; Ren and Candès, 2020). The primary obstacle for incorporating these conditions into DART lies in ensuring that higher layer p-values or statistics also adhere to them, thereby facilitating higher layer FDR control proof via deduction. This intriguing area of research warrants further exploration. Additionally, we aspire to alleviate Condition 4 without relying on the presumption that the distance matrix primarily represents co-status patterns. Our objective is to devise a robust testing algorithm that ensures FDR control in the absence of this assumption while enhancing power when the assumption is valid.

In conclusion, our paper initiates an attempt at joint AI-statistics modeling to generate data-adaptive, powerful, and high-interpretable analysis results. It can be easily extended to the case where other information implies the co-importance pattern of the features. Such information could from domain knowledge, external data sets, or other resources. In addition, the hierarchical testing ideas and techniques can also be extended to solve other multiple testing problems.

8. Acknowledgment

The microbiome samples were collected and sequenced at Memorial Sloan Kettering Cancer Center (MSKCC) and pre-processed at Duke Cancer Institute (DCI) Bioinformatics Shared Resource (BSR). We thank Tsoni Peled and Marcel van den Brink from MSKCC for their help in sample collection and sequencing. We thank Kouros Owzar and Alexander Sibley from DCI-BSR for the help in data pre-processing and constructive discussions. Xuechan Li's research is supported by Duke University. Anthony Sung's research is supported by NIH Award 1-R01-HL151365. Jichun Xie's research is supported by NIH Award 1-R01-HG012555-01.

Appendix Appendix A. More Definitions

In Section 4, we introduced a few terms, including moderate SNR nodes, strong SNR hypotheses, and weak SNR hypotheses. We provide their mathematical definitions here.

Appendix A.1 Moderate SNR Nodes

For any node A, we define its descendant set as

$$\mathcal{D}(A) = \{D : \exists\ \ell,\ \text{such that } D \in \mathcal{A}^{(\ell)} \text{ and } D \subsetneqq A\}.$$

The descendant set contains all the nodes from previous layers that a subset of A.

Definition 4 (Moderate SNR node) *A node A is called a moderate SNR node if*

$$\mathrm{P}\{T_A < \alpha_m,\ \forall D \in \mathcal{D}(A),\ T_D \geq \bar{\Phi}(m^{r_1-1}\sqrt{\log m})\} \geq C_1 > 0, \tag{5}$$

To provide more intuitions on the moderate SNR nodes, we provide a sufficient condition for a node being a moderate SNR node when the test statistics are Gaussian-distributed.

Example 1 *A sequence of independent Gaussian-distributed test statistics $Z_i \sim \mathrm{N}(\tau_i, 1)$ for $i \in [m]$. To test the two-sided hypothesis*

$$H_i : \tau_i = 0 \quad \textit{versus} \quad \tau_i \neq 0,$$

we derive the P-value $T_i = 2\bar{\Phi}(|Z_i|)$.

In example 1, if a node A satisfies

$$\forall i \in A, \quad |\tau_i| \in \left(\frac{\gamma_m}{\sqrt{|A|}}, \frac{\beta_m}{\sqrt{|A|-1}}\right)$$
$$\text{with } \beta_m = \sqrt{2(1-r_1)\log m - 2\log\log m} \text{ and } \gamma_m = \sqrt{2\log m + \log\log\log m}, \tag{6}$$

then A has moderate SNR. We request each τ_i falls in the range involving β_m and γ_m; both slowly increase with m. In practice, the test statistics are calculated based on the observed samples. We usually consider the sample size n increases with the number of hypotheses m. As sample size n increases, the alternative SNR τ_i will also increase, often at the rate of $\sqrt{n}$. Thus, we usually expect to have some alternatives whose SNR falls into the range.

In (6), each τ_i needs to fall in the range. More generally, the purpose to define moderate SNR nodes is to define a set of alternative nodes (a) remaining in tree till they become candidate nodes, and (b) having large enough signals to be discovered when they become candidate nodes. We only need $O(\log m)$ such nodes to make sure the multiple testing procedure is stable and asymptotic valid.

Appendix A.2 Strong SNR Hypotheses

The definition of strong SNR hypotheses is linked with the definition of strong SNR node. On layer 1, we define strong SNR hypotheses. On layer ℓ with $\ell \geq 2$, recursively, we exclude the strong SNR hypotheses defined on previous layers from each node, and then evaluate if this node is a strong SNR node; if so, all the hypotheses in this node are counted as strong SNR hypotheses. This process ends till reaching the top of the tree. Denote the strong SNR node set on layer ℓ by $\mathcal{G}_{\text{st}}^{(\ell)}$ and the strong SNR hypothesis set on layer ℓ by $\Omega_{\text{st}}^{(\ell)}$. To initiate, let $\Omega_{\text{st}}^{(0)} = \emptyset$.

Definition 5 (Strong SNR hypotheses) *On layer ℓ, after excluding the strong SNR hypotheses from the previous layers, define*

$$\tilde{\mathcal{A}}^{(\ell)} = \{A \setminus \cup_{\ell'=1}^{\ell-1} \Omega_{st}^{(\ell')} : A \in \mathcal{A}^{(\ell)}\}.$$

For any node $\tilde{A} \in \tilde{\mathcal{A}}^{(\ell)}$, if for all $i \in \tilde{A}$,

$$\mathrm{P}\{T_i \in \kappa(|\tilde{A}|)\} > 1 - o(m^{-r_1}) \text{ with } \kappa(|\tilde{A}|) = [m^{-\frac{1-r_1}{|\tilde{A}|-1}}, \left\{m(\log m \log\log m)^{1/2}\right\}^{-1/|\tilde{A}|}], \quad (7)$$

Then $\tilde{A}$ is called a strong SNR node. The strong SNR hypothesis set on layer ℓ is

$$\Omega_{st}^{(\ell)} = \{i \in \tilde{A} : \tilde{A} \text{ is a strong SNR node in } \tilde{\mathcal{A}}^{(\ell)}\}.$$

The overall strong SNR hypothesis set is defined as $\Omega_{st} = \cup_{\ell=1}^{L-1} \Omega_{st}^{(\ell)}$.

With a high probability converging to 1, no hypothesis in $\Omega_{\text{st}}^{(\ell)}$ will be rejected before layer ℓ, but all of them will be rejected on layer ℓ.

Note that on layer $\ell \geq 2$, it is possible that $\left\{m(\log m \log\log m)^{1/2}\right\}^{-1/|\tilde{A}|} < m^{-\frac{1-r_1}{|\tilde{A}|-1}}$, which leads $\kappa(|\tilde{A}|) = \emptyset$. In that case, strong SNR nodes do not exist. Our method does not require the existence of the strong SNR nodes.

Under Example 1, (7) is satisfied when the SNR

$$|\tau_j| \in \left(\frac{\gamma_m}{\sqrt{|\tilde{A}|}} + \lambda_m, \frac{\beta_m}{\sqrt{|\tilde{A}|-1}} - \lambda_m\right) \text{ with } \beta_m, \gamma_m \text{ defined in (6), } \lambda_m = \sqrt{2r_1 \log m}. \quad (8)$$

Appendix A.3 Weak SNR Hypotheses

Weak SNR hypotheses are those with weak SNRs so that they are very unlikely to be rejected if aggregated with other null hypotheses.

Definition 6 (Weak SNR hypothesis) *For any alternative hypothesis $i \in \Omega_1$, if*

$$\mathrm{P}(T_i \in \iota) = o(m^{-r_1}) \text{ with } \iota = (0, m^{\frac{r_1-1}{M^{L-1}}}), \quad (9)$$

this hypothesis is called a weak SNR hypothesis.

Under Example 1, (9) is satisfied if $|\tau_i| \in (0, \beta_m/\sqrt{M^{L-1}})$.

Appendix Appendix B. Proofs of Main Theorems

We introduce some notations before we provide the proofs. On layer ℓ, for a working node $S \in \mathcal{B}^{(\ell)}$, let $\mathcal{U}(S) = \{S' \subset S : S' \in \cup_{\ell'=1}^{\ell-1} \mathcal{B}^{(\ell')}\}$ be the collection of sets in the testing path of S. In addition, let $\mathcal{U}^c(S) = \{S'' \in \cup_{\ell'=1}^{\ell-1} \mathcal{B}^{(\ell')} : S'' \cap S = \emptyset, S'' \cup S \subset A, \text{ for some } A \in \mathcal{A}^{(\ell)}\}$ be the collection of sets that was planning to combined with S on layer ℓ of the static aggregation tree but rejected on previous layers. When $S \in \mathcal{B}^{(1)}$, we set $\mathcal{U}(S) = \mathcal{U}^c(S) = \emptyset$. We define $G_S(c)$ as the complementary CDF conditional on previous testing results. When $\ell = 1$, we have $S = \{i\} \subset \{1, ..., m\}$, and $G_S(c) = P(Z_i \geq c)$ with $Z_1, \ldots, Z_m \overset{iid}{\sim} N(0, 1)$. When $\ell > 1$, the oracle rejection path for set $S \in \mathcal{B}^{(\ell)}$ is recursively defined as

$$\mathcal{Q}_z^{(1:\ell-1)} = \{z : \forall S' \in \mathcal{U}(S), G_{S'}(Z_{S'}) \geq \hat{t}^{(\ell_{S'})}(\alpha), \forall S'' \in \mathcal{U}^c(S), G_{S''}(Z_{S''}) \leq \hat{t}^{(\ell_{S''})}(\alpha)\},$$

where

$$G_S(c) = \mathrm{P}\left(Z_S \geq c \middle| \mathcal{Q}_z^{(1:\ell-1)}\right)$$

and $Z_S = \sum_{i \in S} Z_i / \sqrt{|S|}$, and $\ell_{S'}, \ell_{S''} \in \{1, ..., \ell - 1\}$ is the value s.t. $S' \in \mathcal{B}^{(\ell_{S'})}$ and $S'' \in \mathcal{B}^{(\ell_{S''})}$, respectively.

Given $Z_1, \ldots, Z_m$ are mutually independent, we have

$$G_S(c) = \mathrm{P}\left(Z_S \geq c \middle| \forall S' \in \mathcal{U}(S), G_{S'}(Z_{S'}) \geq \hat{t}^{(\ell_{S'})}(\alpha)\right)$$

Given the definition of $G_S(c)$, we define the rejection path as

$$\mathcal{Q}^{(1:\ell-1)} = \{x : \forall S' \in \mathcal{U}(S), G_{S'}(X_{S'}) \geq \hat{t}^{(\ell_{S'})}(\alpha), \forall S'' \in \mathcal{U}^c(S), G_{S''}(X_{S''}) \leq \hat{t}^{(\ell_{S''})}(\alpha)\} \quad (10)$$

In addition, for two sequence of real numbers a_m and b_m, we write $a_m = o(b_m)$ when $a_m/b_m \to 0$, and $a_m = O(b_m)$ when $\lim_{m\to\infty} |a_m/b_m| \leq C$ for some constant C. To prove the asymptotic properties of DART, we need the following lemmas. The proofs of lemmas are shown in supplementary materials.

Lemma 7 *Let $\mathcal{P}_i = \{p \in [0,1] : \mathrm{P}(\tilde{T}_i < p) \geq \epsilon(m)\}$ and $\mathcal{P}_i' = \{p \in [0,1] : \mathrm{P}(\tilde{T}_i < p) \geq \epsilon(m)\epsilon'(m)\}$, with $\epsilon(m), \epsilon'(m) \to 0$. For any set of independent random variable $\hat{T}_i \in [0,1]$, and a collection $\mathcal{M} = \{S \subset \{1, ..., m\} : |S| < c_0\}$ with some constant c_0,*
(1) If $\max_{i\in\mathcal{M}} \sup_{p\in\mathcal{P}_i'} \left|\mathrm{P}(\hat{T}_i < p)/\mathrm{P}(\tilde{T}_i < p) - 1\right| \to 0$, *then,*

$$\sup_{S_0\in\mathcal{M}} \sup_{p\geq\epsilon(m)} \left| \frac{\mathrm{P}(\sum_{i\in S_0} \hat{X}_i > c_{S_0}(p))}{\mathrm{P}(\sum_{i\in S_0} \tilde{X}_i > c_{S_0}(p))} - 1 \right| \to 0,$$

(2) If $\lim_{m\to\infty} \max_{i\in\mathcal{M}} \sup_{p\in\mathcal{P}_i'} \left(P(\hat{T}_i < p)/\mathrm{P}(\tilde{T}_i < p) - 1\right) \leq 0$, *then,*

$$\lim_{m\to\infty} \sup_{S_0\in\mathcal{M}} \sup_{p\geq\epsilon(m)} \left(\frac{\mathrm{P}(\sum_{i\in S_0} \hat{X}_i > c_{S_0}(p))}{\mathrm{P}(\sum_{i\in S_0} \tilde{X}_i > c_{S_0}(p))} - 1 \right) \leq 0$$

Here, $\hat{X}_i = \bar{\Phi}^{-1}(\hat{T}_i)$, $\tilde{X}_i = \bar{\Phi}^{-1}(\tilde{T}_i)$ and $c_{S_0}(p)$ is the value s.t. $\mathrm{P}[\sum_{i\in S_0} \tilde{X}_i > c_{S_0}(p)] = p$.

Lemma 8 *Let* $\tilde{\Omega}_0 = \{i : \tilde{T}_i \text{ follows } \mathrm{Unif}(0,1)\}$, $\mathcal{B}_{0a}^{(\ell)} := \{S \in \mathcal{B}_0^{(\ell)} : \exists A \in \mathcal{A}^{(L)} \setminus \mathcal{A}', s.t. S \subset A\}$, *and* $\mathcal{B}_{0b}^{(\ell)} := \{S \in \mathcal{B}_{0a}^{(\ell)} : S \in \tilde{\Omega}_0\}$, *we have:*

(1) $$\max_{S \in \mathcal{B}_{0a}^{(\ell)}} \sup_{c \in [0, \gamma_m]} \left| \frac{G_S(c)}{\bar{\Phi}(c)} - 1 \right| \to 0$$

(2) $$\max_{S \in \mathcal{B}_{0b}^{(\ell)}} \sup_{c \in [0, \bar{\Phi}^{-1}(1/m)]} \left| \frac{\mathrm{P}(X_S > c | \mathcal{Q}^{(1:\ell-1)})}{\mathrm{P}(X_S > c)} - 1 \right| \to 0$$

Lemma 9 *Define*

$$\mathcal{X}^{(\ell)} = \left\{ x : \sum_{S \in \mathcal{B}_0^{(\ell)}} |S| I(T_S < \hat{t}^{(\ell)}) - \sum_{S \in \mathcal{B}_0^{(\ell)}} |S| \hat{t}^{(\ell)} \le \left\{ \sum_{S \in \mathcal{B}_0^{(\ell)}} |S| \hat{t}^{(\ell)} \right\} \epsilon \right\} \tag{11}$$

$$\mathcal{X}'^{(\ell)} = \left\{ x : \left| \frac{\sum_{S \in \mathcal{B}_0^{(\ell)}} |S| I(T_S < \hat{t}^{(\ell)})}{\sum_{S \in \mathcal{B}_0^{(\ell)}} |S| \hat{t}^{(\ell)}} - 1 \right| \ge \epsilon \right\}$$

Then, $\forall \ell = 1, ..., L$, *when the FDR control holds on layer* $1, ..., \ell - 1$,
(1) For all $\epsilon \in (0, \alpha)$, *if* $\mathrm{P}(m\hat{t}^{(\ell)} \ge C c_{md}) \to 1$, *then* $\mathrm{P}(\mathcal{X}^{(\ell)}) = 1 - o(1)$. *Together with* $\lim_{m \to \infty} |\tilde{\Omega}_0|/m = 1$, *we have* $P(\mathcal{X}'^{(\ell)}) = 1 - o(1)$.
(2) On $\cap_{h=1}^{\ell} \mathcal{X}^{(h)}$, *there exist a constant* C *s.t.* $\hat{t}^{(\ell)} \le C m^{r_1 - 1}$.
(3) Let $\hat{c}_S$ *be the rejection threshold for the test node* $S \in \mathcal{B}^{(\ell)}$, *s.t.* $\bar{G}_S(\hat{c}_S) = \hat{t}^{(\ell)}$. *Then on* $\cap_{h=1}^{\ell} \mathcal{X}^{(h)}$,

$$\hat{c}_S > \beta_m, \ \forall S \in \mathcal{B}^{(\ell)},$$

and on $\cap_{h=1}^{\ell-1} \mathcal{X}^{(h)}$,

$$\hat{c}_S < \gamma_m, \ \forall S \in \mathcal{B}^{(\ell)}.$$

Lemma 10

$$\frac{\sum_{S \in \mathcal{B}_0^{(\ell)}} |S| \hat{t}^{(\ell)}}{\sum_{S \in \mathcal{B}^{(\ell)}} |S| I(T_S < \hat{t}^{(\ell)})} = \alpha(1 + o(1)) \tag{12}$$

Proof [Proof of Theorem 3] Let $\mathcal{V}^{(\ell)} = \{S \in \mathcal{B}_0^{(\ell)} : S \subset \mathcal{R}^{(\ell)}\}$ and $\mathcal{W}^{(\ell)} = \{S \in \mathcal{B}_1^{(\ell)} : S \subset \mathcal{R}^{(\ell)}\}$ be the false rejection node set and the rejection node set on layer ℓ, respectively. Define

$$\mathcal{X}_1 = \{S \in \cup_{\ell=2}^{L} \mathcal{W}^{(\ell)} : S \cap \Omega_{\mathrm{st}} \neq \emptyset \text{ and } S \cap \Omega_0 \neq \emptyset\}$$
$$\mathcal{X}_2 = \{S \in \cup_{\ell=2}^{L} \mathcal{W}^{(\ell)} : S \cap \Omega_{\mathrm{wk}} \neq \emptyset, \ S \setminus (\Omega_0 \cup \Omega_{\mathrm{wk}}) = \emptyset \text{ and } S \cap \Omega_0 \neq \emptyset\}$$
$$\mathcal{X}_3 = \{S \in \cup_{\ell=2}^{L} \mathcal{W}^{(\ell)} : S \cap \Omega_1 \setminus (\Omega_{\mathrm{wk}} \cup \Omega_{\mathrm{st}}) \neq \emptyset \text{ and } S \cap \Omega_0 \neq \emptyset\}$$

Then,

$$\begin{aligned}
&\mathrm{P}(\mathcal{X}_1 \neq \emptyset) \leq \mathrm{P}(\mathcal{X}_1 \neq \emptyset | \cap_{\ell=1}^{L} \mathcal{X}^{(\ell)})\, \mathrm{P}(\cap_{\ell=1}^{L} \mathcal{X}^{(\ell)}) + \mathrm{P}((\cap_{\ell=1}^{L} \mathcal{X}^{(\ell)})^c) \leq C m^{r_1} o(m^{-r_1}) + o(1) \to 0 \\
&\mathrm{P}(\mathcal{X}_2 \neq \emptyset) \leq \mathrm{P}(\mathcal{X}_2 \neq \emptyset | \cap_{\ell=1}^{L} \mathcal{X}^{(\ell)})\, \mathrm{P}(\cap_{\ell=1}^{L} \mathcal{X}^{(\ell)}) + \mathrm{P}((\cap_{\ell=1}^{L} \mathcal{X}^{(\ell)})^c) \\
&\qquad \overset{(a)}{\leq} C m^{r_1}\, \mathrm{P}\left[X_S \geq \beta_m \middle| S \in \Omega_{\mathrm{wk}} \cup \Omega_0 \right] + o(1) \\
&\qquad \leq C m^{r_1} o(m^{-r_1}) + o(1) \to 0
\end{aligned}$$

Here, the inequality (a) is based on Lemma 9 (1) and (3). By condition 4, $|\mathcal{X}_3| = o(c_{\mathrm{md}})$, accordingly,

$$\begin{aligned}
\mathrm{P}\left(FDP > \alpha + \epsilon\right) \leq & \mathrm{P}(\mathcal{X}_1 \cup \mathcal{X}_2 \neq \emptyset) + \mathrm{P}\left(\frac{\sum_{\ell=1}^{L} \sum_{S \in \mathcal{V}^{(\ell)}} |S|}{\sum_{\ell=1}^{L} \sum_{S \in \mathcal{R}_{\mathrm{node}}^{(\ell)}} |S|} > \alpha + \epsilon, \mathcal{X}_1 \cup \mathcal{X}_2 = \emptyset \right) \\
\leq & o(1) + \sum_{\ell=1}^{L} \mathrm{P}\left(\frac{\sum_{S \in \mathcal{V}^{(\ell)} \setminus \mathcal{X}_3} |S|}{\sum_{S \in \mathcal{R}_{\mathrm{node}}^{(\ell)}} |S|} > \alpha + \epsilon + o(1) \right) \to 0
\end{aligned}$$

So statement (a) is proved. The statement (b) can be proved in the similar way. ■

Appendix Appendix C. Additional Details on Simulation and Real Data Analysis

All the experiments were conducted on 2.10 GHz Intel Xeon Gold 6252 processors with 16 Gb memory at the Duke Compute Cluster. We requested 80 cores when running the simulation experiments to save time. Experiment code can be found in `https://github.com/jichunxie/DART_manu_support.git`. We also built an R package, which can be found in `https://github.com/jichunxie/DART.git`.

Appendix C.1 Details on Numerical Experiments

We generated four simulation settings, each with $n = 300$ observations on $m = 1000$ features (hypotheses). All four simulation settings were based on a set of parameters $\eta_i, i \in [m]$ related to alternative hypothesis signal levels. We defined two driver features 7 and 156; the features close to them were likely alternative. We also added 10 stand-alone features. Define the stand-alone feature set $\Omega_2 = \{100, 200, \ldots, 1000\}$.

$$\begin{aligned}
&\eta_i' = \{[3.4\phi_1(d_{156,i}) - 0.8] \vee 0\} + 3\{\phi_2(d_{7,i})\} + 10 * I(i \in \Omega_2), \\
&\eta_i = \eta_i' I(\eta_i' > 0.15)
\end{aligned}$$

Here, ϕ_1 and ϕ_2 are the probability density functions of $\mathrm{N}(0, 0.8)$ and $\mathrm{N}(0, 0.05)$, respectively. Once feature locations and signals η_is were generated, they were fixed across all settings and all repetitions. We visualized the feature locations and their η_i in Figure 4.

Below is the list of the four settings.

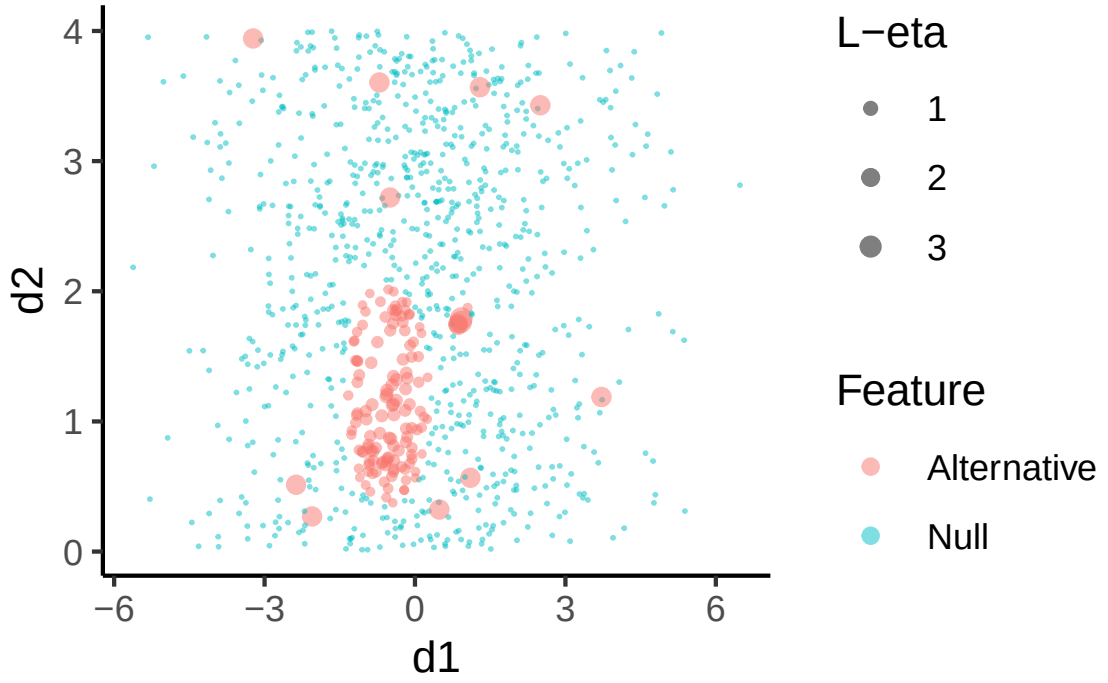


Figure 4: Illustration of the simulated hypotheses' affiliated location and their corresponding η_i. A dot stands for a hypothesis. The dot color indicates its corresponding hypothesis status; and its size is L-eta $= \log(\eta_i + 1) + 0.01$.

SE1: For node $i \in \{1, ..., m\}$, we generated the P-values $T_i = 2\bar{\Phi}(|\breve{Z}_i|)$ with $\breve{Z}_1, \ldots, \breve{Z}_m$ independently from $\mathrm{N}(\sqrt{n}\theta_i, 1)$, with $\theta_i = \frac{1}{5}\eta_i$.

SE2: For node $i \in \{1, ..., m\}$, we generated the P-values $T_i = 2\bar{\Phi}(|\breve{Z}_i|)$ with $\breve{Z}_1, \ldots, \breve{Z}_m$ independently from the mixed Gaussian and T distribution $0.04\mathrm{t}_5(\sqrt{n}\theta_i) + 0.96\mathrm{N}(\sqrt{n}\theta_i, 1)$. Here, $\mathrm{t}_5(\sqrt{n}\theta_i)$ stands for the student t distribution with the degree of freedom 5 and none centrality parameter $\sqrt{n}\theta_i$, with $\theta_i = \frac{1}{5}\eta_i$.

SE3: Consider the linear mode $Y_i = \vartheta_{0,i} + \theta_i W_1 + \vartheta_{2,i} W_2 + \epsilon_i$. Here, W_1 and W_2 were generated from Binom(0.5) and Unif(0.1, 0.5); ϵ_i was from $N(0, 1)$. We set $\vartheta_{0,i} = \vartheta_{2,i} = 0.1$ and $\theta_i = \frac{2}{5}\eta_i$. The P-values T_i were generated from the Wald test of testing whether θ_i is zero.

SE4: We generated the data $\mathcal{D} = \{(W_{1,j}, W_{2,j}, \Delta_{ij}, E_{ij}) : i \in [m], j \in [n]\}$. Here, the covariates $W_{1,j}$ and $W_{2,j}$ were sampled from Binom(0.5) and Unif(0.1, 0.5), respectively. $E_{ij} \in [0, +\infty)$ is the survival time and $\Delta_{ij} \in \{0, 1\}$ is the event indicator. The true event time was generated from the exponential distribution with the rate $\exp\{\theta_i W_{1,j} + \vartheta_i W_{2,j}\}$, where $\vartheta_i = 0.1$ and $\theta_i = \frac{1}{2}\eta_i$. The centering time C_{ij} was generated from Unif(0, 5). The observed event time was set as $E_{ij} = \min\{\tilde{E}_{ij}, C_{ij}\}$. We used the Cox regression model to regress the covariates $W_{1,j}$ and $W_{2,j}$ on the event (Δ_{ij}, E_{ij}) Cox (1972). The P-values T_i were generated from the Wald test of testing whether θ_i is zero.

To check the robustness of the algorithms when the distance cannot fully reflect the co-status patterns, we switched some hypotheses' null/alternative statuses. In particular, we randomly changed proportion τ of alternative hypotheses to be null and proportional τ of the null hypotheses to be alternative. The switching rate τ reflects the violation level of the co-status patterns. Here is the detailed switching process.

(1) Denote the original null and alternative sets by $\Omega_0' = \{i : \theta_i = 0\}$ and $\Omega_1' = \{i : \theta_i > 0\}$.

(2) We randomly selected the elements from the original null and alternative sets to form

- the alternative-to-null set $\Omega_{1,0}' \subset \Omega_1'$, where $|\Omega_{1,0}'| = \lfloor \tau |\Omega_1'| \rfloor$; and
- the null-to-alternative set $\Omega_{0,1}' \subset \Omega_0'$, where $|\Omega_{0,1}'| = |\Omega_{1,0}'|$

(3) Re-assign the signal parameters.

- For $i \in \Omega_{0,1}'$, re-assign $\theta_i = \theta_j$ for a random-chosen $j \in \Omega_{1,0}'$.
- For $i \in \Omega_{1,0}'$, set $\theta_i = 0$.

(4) Define the new null and alternative sets $\Omega_0 = \Omega_0' \cup \Omega_{1,0}' \setminus \Omega_{0,1}'$ and $\Omega_1 = \Omega_1' \cup \Omega_{0,1}' \setminus \Omega_{1,0}'$.

The average FDP and sensitivity across 200 repetitions were summarized in Figure 2B and Figure 5.

We applied DART, BH (Benjamini and Hochberg, 1995), two FDR_L procedures (Zhang et al., 2011), and AdaPT (Lei and Fithian, 2018) to the above numerical settings. BH does not have tuning parameters. For DART, based on the tuning parameter selection criterion in Section 3.4, we set $M = 3$ and constructed a 3-layer aggregation tree, with distance thresholds $g^{(2)} = 0.88$ and $g^{(3)} = 1.52$. For the two FDR_L procedures, Zhang et al. (2011) recommended setting k as an odd number greater than three and used $k = 5$ in the simulation experiments. Thus, we set $k = 5$ in our numerical experiments too. For AdaPT, we followed the instructions found at `https://cran.r-project.org/web/packages/adaptMT/vignettes/adapt_demo.html` to set up its tuning parameters. During simulation, we noticed that AdaPT sometimes failed to generate results after a long execution time. To ensure valid results from AdaPT, Figures 2 and 5 only summarize the repetition who successfully deliver a result. Table 1 lists the number of repetitions that failed under different scenarios among 200 repetitions.

SE	$\tau = 0$	$\tau = 0.02$	$\tau = 0.04$	$\tau = 0.06$	$\tau = 0.08$	$\tau = 0.1$
1	33	41	34	45	46	45
2	34	37	29	42	39	45
3	23	30	36	42	32	51
4	12	19	15	27	25	36

Table 1: Number of repetition fails to deliver testing result within 8 minutes.

Appendix C.2 Experiments to Decide the Optimal Maximum Node Size

We considered four settings of the maximum node sizes: $M \in \{2, 3, 4, 5\}$. To design a good tuning parameter selection criterion, we evaluated DART's performance with different M via numerical experiments described in Section 5. The maximum number of layers L and distance thresholds $g^{(\ell)}$ were set according to Section 3.4 and Algorithm 3. We listed the resulting tuning parameters in Table 2.

Figure 6 shows that no matter what M we used, DART always controls the average FDR well under SE1, SE3, and SE4. For SE2, when the nominal FDR is 5%, DART has slight inflation because SE2 deliberately misspecified the null distribution. However, when the nominal FDR $\geq 10\%$, DART under SE2 still has the average FDP well controlled. This indicates that DART's performance is robust in M. When $M \in \{2, 3\}$, DART has higher sensitivity. Thus, in practice, we recommend using $M = 2$ or $M = 3$.

M	2	3	4	5
Maximum Layer L	4	3	2	2
Distance thresholds	$g^{(2)} = 1.20$ $g^{(3)} = 1.52$ $g^{(4)} = 1.74$	$g^{(2)} = 0.88$ $g^{(3)} = 1.52$	$g^{(2)} = 0.25$	$g^{(2)} = 0.19$

Table 2: Summary of the tuning parameters selected under different M

Appendix C.3 Aggregation Tree Construction in Real-world Experiment

Because 9 ASVs are chosen as the reference ASVs, the distance matrix is calculated among the remaining 857 non-reference ASV using the R package `Phangorn` (Schliep, 2011) based on the JC69 model (Jukes et al., 1969). As the default model in `Phangorn`, the JC69 model is a classical Markov model of DNA sequence evolution and can be used to estimate the evolutionary distance between sequences. Two ASVs with similar sequences tend to be evolutionary close to each other, and more likely to perform similar biological functions. Therefore, we incorporate the distance matrix in identifying the important ASVs.

Based on the tuning parameter selection procedure described in Section 2.3, we construct an aggregation tree with $M = 3$, $L = 3$. The set of possible threshold G is set as $\{4, 8, 12, 16\}/\sqrt{n \log m \log\log m}$, with $n = 456$ and $m = 857$, and we choose $g^{(2)} = g^{(3)} = 16/\sqrt{n \log m \log\log m} = 0.21$.

Algorithm 3: $g^{(\ell)}$ Selection algorithm.

Data: Distance Matrix $D = (d_{ij})_{m\times m}$, Sample size n, number of features m, the maximum children number M, the maximum layer L

Result: $g^{(2)}, ..., g^{(L)}$.

// set searching upper bound $d_{\max}$ and step-size $s_{n,m}$

Let $d_{\max} = \max_{j\in\Omega} \min_{i\in\{i:i\neq j\}} d_{ij}$; $s_{n,m} = 4/\sqrt{n\log(m)\log\log(m)}$;

for $\ell = 2, ..., L$ **do**

// on layer ℓ, search $g^{(\ell)}$ from $(g^{(\ell-1)}, d_{\max}]$, $g^{(1)} = 0$

Let M_g =NULL; e_g=1; G =NULL; $g = g^{(\ell-1)} + s_{n,m}$;

while $g \leq (2M^{L-2} - 1)d_{\max}$ *and* $e_g < 10$ **do**

// stop searching process if the value g exceed the searching upper bound or the $|\tilde{A}^{(\ell)}(g)|$ does not increase for past 10 candidate values g.

// stop searching process if the value g exceed the searching upper bound.

Use Algorithm 1 to Construct an ℓ layers aggregation tree $\mathcal{T}_\ell = \{\mathcal{A}^{(\ell')} : \ell' = 1, ..., \ell\}$ with maximum children number M, and $(g^{(1)}, ..., g^{(\ell-1)}, g)$;

Set $\tilde{A}^{(\ell)}(g) = \{A : A \in \mathcal{A}^{(\ell)}(g),\ |\mathcal{C}(A)| \geq 2\}$; **if** $m_g \geq |\tilde{A}^{(\ell)}(g)|$ **then**

$e_g = e_g + 1$;

else

e_g=1;

$G = (G, g)$; $M_g = (M_g, |\tilde{A}^{(\ell)}(g)|)$; $m_g = |\tilde{A}^{(\ell)}(g)|$;

$g = g + s_{n,m}$;

$g^{(\ell)} = \min\{\arg\max_{g\in G} M_g\}$;

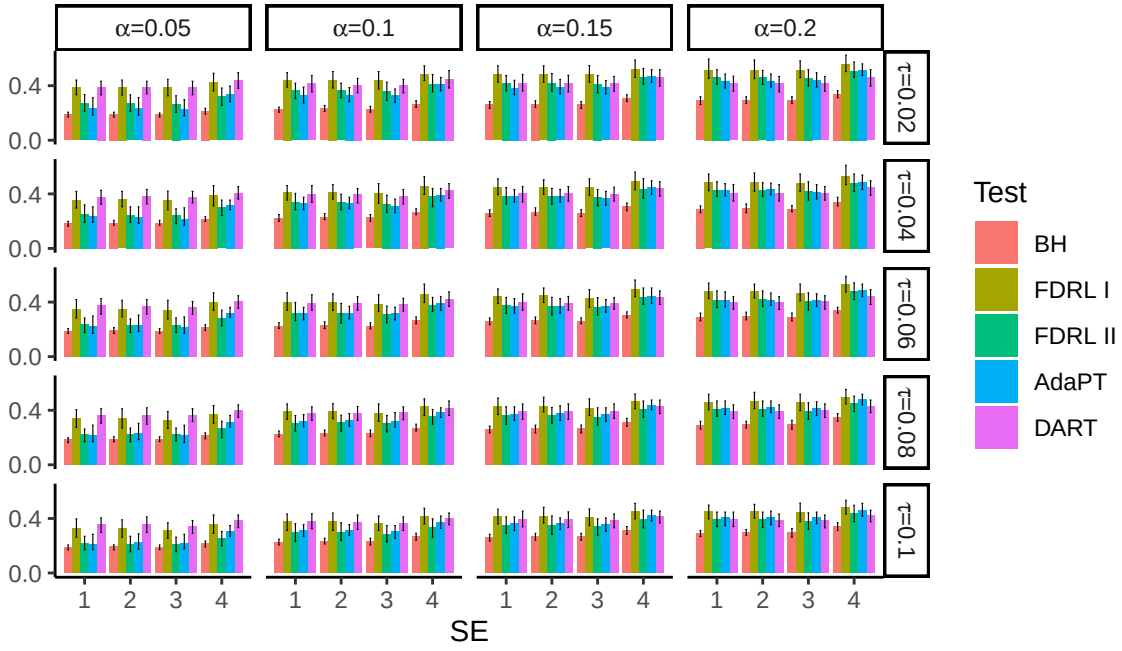


Figure 5: Sensitivity of various testing methods under SE1-SE4 when proportion τ of the alternative hypothesis were switched with the null. The main bars mark the average values across 200 repetitions; the error bars mark their 25% and 75% quantiles. Every column shows the sensitivity under a nominal FDR level α.

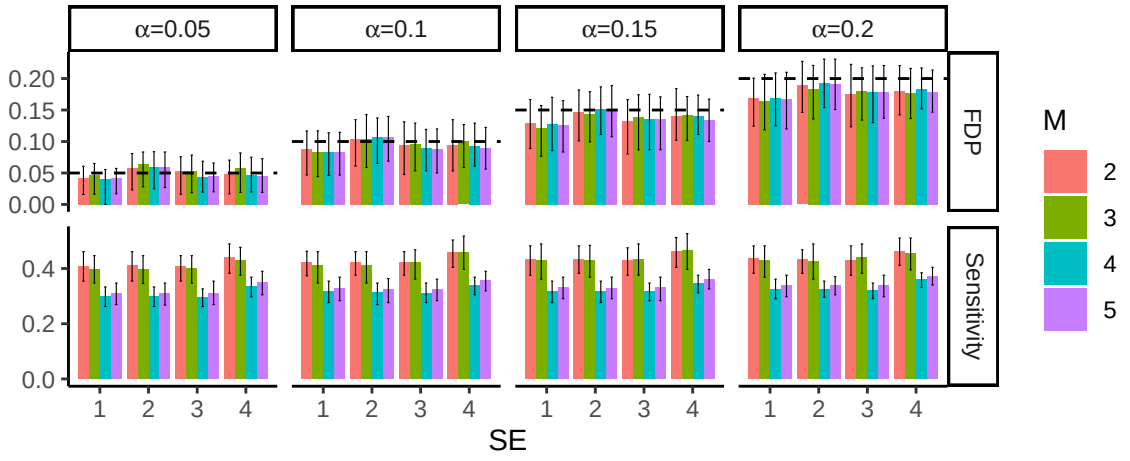


Figure 6: Performance of DART across different M. The main bars mark the average values across 200 repetitions; the error bars mark their 25% and 75% quantiles. Every column shows the performance under a nominal FDR level α. The first row represents the average FDP, and the dashed horizontal lines marks the desired FDR level. The second row represents the average sensitivity.

References

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B*, 44(2):139–160, 1982.

Aaron F Alexander-Bloch, Petra E Vértes, Reva Stidd, François Lalonde, Liv Clasen, Judith Rapoport, Jay Giedd, Edward T Bullmore, and Nitin Gogtay. The anatomical distance of functional connections predicts brain network topology in health and schizophrenia. *Cereb Cortex*, 23(1):127–38, Jan 2013. doi: 10.1093/cercor/bhr388.

Denis Belomestny and Vladimir Spokoiny. Spatial aggregation of local likelihood estimates with applications to classification. *The Annals of Statistics*, 35(5):2287–2311, 2007. ISSN 00905364. URL `http://www.jstor.org/stable/25464582`.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. doi: 10.1088/1742-5468/2008/10/p10008. URL `https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008`.

T Tony Cai and Weidong Liu. Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, 111(513):229–240, 2016a. doi: 10.1080/01621459.2014.999157.

T Tony Cai and Weidong Liu. Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, 111(513):229–240, 2016b.

T Tony Cai, Wenguang Sun, and Yin Xia. Laws: A locally adaptive weighting and screening approach to spatial multiple testing. *Journal of the American Statistical Association*, pages 1–30, 2020.

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2: high-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581, 2016.

Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–13, Aug 2012. doi: 10.1093/bioinformatics/bts342.

Marcus J Claesson, Ian B Jeffery, Susana Conde, Susan E Power, Eibhlís M O'connor, Siobhán Cusack, Hugh MB Harris, Mairead Coakley, Bhuvaneswari Lakshminarayanan, Orla O'Sullivan, et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410):178–184, 2012.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972.

Alex Dmitrienko and Ajit C Tamhane. General theory of mixture procedures for gatekeeping. *Biometrical Journal*, 55(3):402–19, May 2013. doi: 10.1002/bimj.201100258.

Alex Dmitrienko, Brian Wiens, Ajit Tamhane, and Xin Wang. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered objectives. *Statistics in Medicine*, 26: 2465–78, 05 2007. doi: 10.1002/sim.2716.

Alex Dmitrienko, Ajit C Tamhane, Lingyun Liu, and Brian L Wiens. A note on tree gatekeeping procedures in clinical trials. *Statistics in Medicine*, 27(17):3446–51, Jul 2008. doi: 10.1002/sim.3307.

Alex Dmitrienko, George Kordzakhia, and Ajit C Tamhane. Multistage and mixture parallel gatekeeping procedures in clinical trials. *Journal of Biopharmaceutical Statistics*, 21(4): 726–47, Jul 2011. doi: 10.1080/10543406.2011.551333.

Anders Eklund, Mats Andersson, Camilla Josephson, Magnus Johannesson, and Hans Knutsson. Does parametric fmri analysis with spm yield valid results? an empirical study of 1484 rest datasets. *Neuroimage*, 61(3):565–78, Jul 2012. doi: 10.1016/j.neuroimage.2012.03.093.

Ronald Fisher. *Statistical method for research workers*. Oliver and Boyd, Edinburgh ;London, 1925.

Tanya P Garcia, Samuel Müller, Raymond J Carroll, and Rosemary L Walzem. Identification of important regressor groups, subgroups and individuals via regularization methods:

application to gut microbiome data. *Bioinformatics*, 30(6):831–7, Mar 2014. doi: 10.1093/bioinformatics/btt608.

Jelle J Goeman and Livio Finos. The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical Applications in Genetics and Molecular Biology*, 11(1): Article 11, Jan 2012. doi: 10.1515/1544-6115.1554.

Robert R Jenq, Carles Ubeda, Ying Taur, Clarissa C Menezes, Raya Khanin, Jarrod A Dudakov, Chen Liu, Mallory L West, Natalie V Singer, Michele J Equinda, et al. Regulation of intestinal inflammation by microbiota following allogeneic bone marrow transplantation. *Journal of Experimental Medicine*, 209(5):903–911, 2012.

Xin Jin and Jiawei Han. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_425. URL `https://doi.org/10.1007/978-0-387-30164-8_425`.

Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3:21–132, 1969.

Daniel Kristanto, Mianxin Liu, Xinyang Liu, Werner Sommer, and Changsong Zhou. Predicting reading ability from brain anatomy and function: From areas to connections. *Neuroimage*, 218:116966, 09 2020. doi: 10.1016/j.neuroimage.2020.116966.

Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, 11(5), 2015.

Donghwan Lee and Youngjo Lee. Extended likelihood approach to multiple testing with directional error control under a hidden markov random field model. *Journal of Multivariate Analysis*, 151:1 – 13, 2016. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2016.07.001. URL `http://www.sciencedirect.com/science/article/pii/S0047259X16300458`.

Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B*, 80(4):649–679, 2018.

Lihua Lei, Aaditya Ramdas, and William Fithian. A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, 108(2):253–267, 07 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa064. URL `https://doi.org/10.1093/biomet/asaa064`.

Xiaoqing Li, Yu Lin, Xue Li, Xiaoxiao Xu, Yanmin Zhao, Lin Xu, Yang Gao, Yixue Li, Yamin Tan, Pengxu Qian, et al. Tyrosine supplement ameliorates murine agvhd by modulation of gut microbiome and metabolome. *EBioMedicine*, 61:103048, 2020a.

Yimei Li, John H Gilmore, Dinggang Shen, Martin Styner, Weili Lin, and Hongtu Zhu. Multiscale adaptive generalized estimating equations for longitudinal neuroimaging data. *Neuroimage*, 72:91–105, May 2013. doi: 10.1016/j.neuroimage.2013.01.034.

Yunxiao Li, Yi-Juan Hu, and Glen A. Satten. A bottom-up approach to testing hypotheses that have a branching tree dependence structure, with error rate control. *Journal of the American Statistical Association*, pages 1–18, sep 2020b. doi: 10.1080/01621459.2020.1799811. URL https://doi.org/10.1080%2F01621459.2020.1799811.

Dandan Lin, Bo Hu, Pengfei Li, Ye Zhao, Yang Xu, and Depei Wu. Roles of the intestinal microbiota and microbial metabolites in acute gvhd. *Experimental Hematology & Oncology*, 10(1):1–19, 2021.

Weidong Liu et al. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.

Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020. doi: 10.1080/01621459.2018.1554485.

Jackson H Loper, Lihua Lei, William Fithian, and Wesley Tansey. Smoothed nested testing on directed acyclic graphs. *Biometrika*, 109(2):457–471, 2022.

Jennifer B H Martiny, Stuart E Jones, Jay T Lennon, and Adam C Martiny. Microbiomes in light of traits: A phylogenetic perspective. *Science*, 350(6261):aac9323, Nov 2015. doi: 10.1126/science.aac9323.

Rosa J Meijer and Jelle J Goeman. A multiple testing method for hypotheses structured in a directed acyclic graph. *Biometrical Journal*, 57(1):123–143, 2015.

Mathilde Payen, Ioannis Nicolis, Marie Robin, David Michonneau, Johanne Delannoye, Camille Mayeur, Nathalie Kapel, Béatrice Berçot, Marie-José Butel, Jérôme Le Goff, et al. Functional and phylogenetic alterations in gut microbiome are linked to graft-versus-host disease severity. *Blood Advances*, 4(9):1824–1832, 2020.

Alessio Perinelli, Davide Tabarelli, Carlo Miniussi, and Leonardo Ricci. Dependence of connectivity on geometric distance in brain networks. *Scientific Reports*, 9(1):13412, 09 2019. doi: 10.1038/s41598-019-50106-2.

Aaditya Ramdas, Jianbo Chen, Martin J Wainwright, and Michael I Jordan. A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika*, 106(1): 69–86, 2019a.

Aaditya K. Ramdas, Rina F. Barber, Martin J. Wainwright, and Michael I. Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47(5), oct 2019b. doi: 10.1214/18-aos1765. URL https://doi.org/10.1214%2F18-aos1765.

Zhimei Ren and Emmanuel Candès. Knockoffs with side information. 01 2020. URL https://arxiv.org/pdf/2001.07835.pdf.

K.P. Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011. URL https://doi.org/10.1093/bioinformatics/btq706.

Yusuke Shono, Melissa D Docampo, Jonathan U Peled, Suelen M Perobelli, Enrico Velardi, Jennifer J Tsai, Ann E Slingerland, Odette M Smith, Lauren F Young, Jyotsna Gupta, et al. Increased gvhd-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice. *Science Translational Medicine*, 8(339):339ra71–339ra71, 2016.

Hai Shu, Bin Nan, and Robert Koeppe. Multiple testing for neuroimaging via hidden markov random field. *Biometrics*, 71(3):741–750, 2015.

J. Soriano and L. Ma. Probabilistic multi-resolution scanning for two-sample differences. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 79:547–572, 2017.

Wesley Tansey, Yixin Wang, David Blei, and Raul Rabadan. Black box FDR. In *International conference on machine learning*, pages 4867–4876. PMLR, 2018.

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9, 2019.

Matt Turek. Explainable artifical intelligence (xai). https://www.darpa.mil/program/explainable-artificial-intelligence, 2021.

Dong Xi and Ajit C Tamhane. A general multistage procedure for k-out-of-n gatekeeping. *Statistics in Medicine*, 33(8):1321–35, Apr 2014. doi: 10.1002/sim.6028.

Fei Xia, Martin J Zhang, James Y Zou, and David Tse. Neuralfdr: Learning discovery thresholds from hypothesis features. *Advances in neural information processing systems*, 30, 2017.

Jichun Xie and Ruosha Li. False discovery rate control for high dimensional networks of quantile associations conditioning on covariates. *Journal of the Royal Statistical Society: Series B*, 80(5):1015–1034, Nov 2018. doi: 10.1111/rssb.12288.

Daniel Yekutieli. Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008. ISSN 01621459. URL `http://www.jstor.org/stable/27640041`.

Marie Lisandra Zepeda-Mendoza and Osbaldo Resendis-Antonio. *Hierarchical Agglomerative Clustering*, pages 886–887. Springer New York, New York, NY, 2013. ISBN 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_1371. URL `https://doi.org/10.1007/978-1-4419-9863-7_1371`.

Chunming Zhang, Jianqing Fan, and Tao Yu. Multiple testing via FDR_L for large scale imaging data. *Annals of Statistics*, 39(1):613, 2011.

Supplementary Materials for "DART: Distance Assisted Recursive Testing"

Appendix S1. Proofs of Lemmas

Proof [Proof of Lemma 1] Since the proof of the theorem statement (b) is similar to the proof of the theorem statement (a), we will only focusing on the proof of statement (a).

The random variable $FDP^{(\ell)}$ can be decomposed to the product of two parts.

$$FDP^{(\ell)} = \frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I\{T_S < \hat{t}^{(\ell)}\}}{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}} \times \frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}}{\max(\sum_{S\in\mathcal{B}^{(\ell)}}|S|I\{T_S < \hat{t}^{(\ell)}\}, 1)} \tag{S1}$$

Based on (S1), in order to prove $\lim_{m\to\infty}\mathrm{P}(FDP^{(\ell)} \le \alpha+\epsilon) = 1$ for all $\epsilon > 0$, we only need prove

$$\lim_{m\to\infty}\mathrm{P}\left\{\frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I\{T_S < \hat{t}^{(\ell)}\}}{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}} - 1 < \epsilon\right\} \to 1 \tag{S2}$$

$$\lim_{m\to\infty}\mathrm{P}\left\{\left|\frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}}{\max(\sum_{S\in\mathcal{B}^{(\ell)}}|S|I\{T_S < \hat{t}^{(\ell)}\}, 1)} - \alpha\right| > \epsilon\right\} \to 0 \tag{S3}$$

(S3) is immediately followed by Lemma 10, and we will prove (S2) by induction. Below is a list of the proof sketch:

1. On layer 1, show $P(m\hat{t}^{(1)} \ge Cc_{\mathrm{md}}) \to 1$. Then, by applying Lemma 9, we have
   - $P(\mathcal{X}^{(1)}) \to 1$, which is equivalent to (S2). Hence, we proved the FDR control on layer 1.
   - $P(\beta_m < \hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(1)}) \to 1$, and $P(\hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(2)}) \to 1$. Note that although this conclusion is not used to prove the FDR control on the current layer, but is necessary to guarantee the FDR control on higher layers.
2. On layer $\ell \ge 2$, assume the FDR control holds on previous layers and $P(\mathcal{X}^{(\ell')}) \to 1$ for all $\ell' = 1, \ldots, \ell-1$. Then by Lemma 9, $P(\beta_m < \hat{c}_S < \gamma_m, \forall S \in \cup_{\ell'=1}^{\ell-1}\mathcal{B}^{(\ell')}) \to 1$, and $P(\hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(\ell)}) \to 1$. Accordingly, we can get $P(m\hat{t}^{(1)} \ge Cc_{\mathrm{md}}) \to 1$. Then, by applying the Lemma 9 again, we have
   - $P(\mathcal{X}^{(\ell)}) \to 1$, which is equivalent to (S2). Hence, we proved the FDR control on layer ℓ.
   - $P(\beta_m < \hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(\ell)}) \to 1$, and $P(\hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(\ell+1)}) \to 1$.

We start the proof on layer 1. **Layer 1:**

Take a subset $\mathcal{F}^{(1)} \subset \mathcal{A}_{\mathrm{md}} \cap \mathcal{A}^{(1)}$, such that $|\mathcal{F}^{(1)}| = c_{\mathrm{md}}$. For any $i \in \mathcal{F}^{(1)}$, we have $\mathrm{P}(X_i > \gamma_m) \ge C$. By Markov's inequality, we have:

$$\mathrm{P}\left(\Big|\sum_{i\in\mathcal{F}^{(1)}} I(X_i > \gamma_m) - \sum_{i\in\mathcal{F}^{(1)}}\mathrm{P}(X_i > \gamma_m)\Big| \ge c_{\mathrm{md}}^{3/4}\right) \le C(c_{\mathrm{md}})^{-1/2}$$

Thus,

$$\mathrm{P}\big[\sum_{1\le i\le m} I(T_i \le \hat{t}^{(1)}) \ge Cc_{\mathrm{md}} - c_{\mathrm{md}}^{3/4}\big] \ge 1 - o(1)$$

Therefore, by Lemma 10, exists constant $C^{(1)}$, s.t.

$$\mathrm{P}\left[m_0\hat{t}^{(1)} \ge C^{(1)}c_{\mathrm{md}}\right] \ge 1 - o(1) \tag{S4}$$

Together with Lemma 9 (1), we have $\mathrm{P}(\mathcal{X}^{(1)}) \to 1$ and accordingly, $\mathrm{P}(FDP^{(1)} < \alpha + \epsilon) \to 1$.

Layer ℓ:

Based on similar arguments on Layer 1, it suffices to show $\mathrm{P}(m_0\hat{t}^{(\ell)} > C^{(\ell)}c_{\mathrm{md}}) \to 1$ for some constant $C^{(\ell)}$.

Assume $\forall h = 1, \ldots, \ell - 1$, $\mathrm{P}(\mathcal{X}^{(h)}) \to 1$, then by Lemma 9, we have $\mathrm{P}(\beta_m < \hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(h)}) \to 1$, and $\mathrm{P}(c_S < \gamma_m, \forall S \in \mathcal{B}^{(\ell)}) \to 1$.

Let $\mathcal{F}^{(\ell)} \subset \mathcal{A}_{\mathrm{md}} \cap \mathcal{A}^{(\ell)}$ with $|\mathcal{F}^{(\ell)}| = c_{\mathrm{md}}$. Define

$$\hat{\mathcal{F}}^{(\ell)} = \left\{A \in \mathcal{B}^{(\ell)} \cap \mathcal{F}^{(\ell)} : T_A < \alpha_m\right\}$$

By condition 2, $\forall A \in \mathcal{F}^{(\ell)}$,

$$\mathrm{P}(A \in \hat{\mathcal{F}}^{(\ell)}) \ge \mathrm{P}\left(T_A < \alpha_m, T_D \ge \bar{\Phi}(m^{r_1-1}\sqrt{\log m}), \forall D \in \mathcal{D}(A)\right) \ge C_1 \tag{S5}$$

Accordingly, define $\hat{\mathcal{X}}^{(\ell)} = \{|\hat{\mathcal{F}}^{(\ell)}| \ge c_{\mathrm{md}}/2\}$, then $\mathrm{P}(\hat{\mathcal{X}}^{(\ell)}) \ge 1 - o(1)$.

On $\hat{\mathcal{X}}^{(\ell)}$, we have

$$\sum_{S\in\mathcal{B}_1^{(\ell)}} I(T_S \le \hat{t}^{(\ell)}) \ge Cc_{\mathrm{md}}$$

Then based on Lemma 8, we can conclude that $\mathrm{P}(m_0\hat{t}^{(\ell)} \ge C^{(\ell)}c_{\mathrm{md}}) \ge 1 - o(1)$ for some constant $C^{(\ell)}$. ■

Proof [Proof of Lemma 7] (1) Define $\tilde{X}_i = \bar{\Phi}(\tilde{T}_i)$, For $k \in \{1, ..., c_0\}$, let $q_0 \ge \epsilon(m)$. Also define $b_{1,k}(q_0)$, c_1,...,c_k be the value s.t. $P(\sum_{j=1}^{k}\tilde{X}_j > b_{1,k}(q_0)) = q_0\big[\epsilon'(m)\big]^{(c_0-k)/c_0}$, and $P(\tilde{X}_1 > c_1) = ... = P(\tilde{X}_k > c_k) = \epsilon(m)\epsilon'(m)$, respectively. For simplicity's sake, we use $b_{1,k}$ to present $b_{1,k}(q_0)$.

Based on the definition, we have

$$b_{1,k} < \sum_{j=1}^{k} c_j$$

Thus, when $k = 2$,

$$\begin{aligned}
&P(\hat{X}_1 + \hat{X}_2 > b_{1,2})\\
=&P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_1 > b_{1,2} - c_2, \hat{X}_2 > b_{1,2} - c_1)\\
&+ P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_1 < b_{1,2} - c_2) + P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_2 < b_{1,2} - c_1)\\
=&P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, c_1 > \hat{X}_1 > b_{1,2} - c_2) + P(\hat{X}_1 > c_1, \hat{X}_2 > b_{1,2} - c_1)\\
&+ P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_1 < b_{1,2} - c_2) + P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_2 < b_{1,2} - c_1)
\end{aligned}$$

Based on construction, the last three terms always smaller than $\epsilon(m)\epsilon'(m)(1+\delta_4(m))$ for $\delta_4(m) := \max_{i\in\Omega}\sup_{p\in\mathcal{P}_i'}\left|P(\hat{T}_i<p)/P(\tilde{T}_i<p)-1\right| \to 0$, and accordingly, we have

$$\begin{aligned}
&P(\hat{X}_1+\hat{X}_2>b_{1,2}, c_1>\hat{X}_1>b_{1,2}-c_2)+P(\hat{X}_1>c_1,\hat{X}_2>b_{1,2}-c_1)\\
\leq&[P(\hat{X}_1+\tilde{X}_2>b_{1,2}, c_1>\hat{X}_1>b_{1,2}-c_2)+P(\hat{X}_1>c_1,\tilde{X}_2>b_{1,2}-c_1)](1+\delta_4(m))\\
\leq&[P(\hat{X}_1+\tilde{X}_2>b_{1,2}, \hat{X}_1>b_{1,2}-c_2,\tilde{X}_2>b_{1,2}-c_1)](1+\delta_4(m))\\
\leq&P(\tilde{X}_1+\tilde{X}_2>b_{1,2}, \tilde{X}_1>b_{1,2}-c_2,\tilde{X}_2>b_{1,2}-c_1)(1+\delta_4(m))^2
\end{aligned}$$

Based on similar arguments, we can also have

$$\begin{aligned}
&P(\hat{X}_1+\hat{X}_2>b_{1,2}, c_1>\hat{X}_1>b_{1,2}-c_2)+P(\hat{X}_1>c_1,\hat{X}_2>b_{1,2}-c_1)\\
\geq&P(\tilde{X}_1+\tilde{X}_2>b_{1,2}, \tilde{X}_1>b_{1,2}-c_2,\tilde{X}_2>b_{1,2}-c_1)(1-\delta_4(m))^2
\end{aligned}$$

Thus,

$$\sup_{q_0\geq\epsilon(m)\left[\epsilon'(m)\right]^{\frac{c_0-2}{c_0}}}\left|\frac{P(\hat{X}_1+\hat{X}_2>b_{1,2})}{P(\tilde{X}_1+\tilde{X}_2>b_{1,2})}-1\right|\to 0$$

Similarly, if $\sup_{q_0\geq\epsilon(m)\left[\epsilon'(m)\right]^{\frac{c_0-k}{c_0}}}\left|\frac{P(\sum_{j=1}^k\hat{X}_j>b_{1,k})}{P(\sum_{j=1}^k\tilde{X}_j>b_{1,k})}-1\right|\to 0$, we can have

$$\sup_{q_0\geq\epsilon(m)\left[\epsilon'(m)\right]^{\frac{c_0-k-1}{c_0}}}\left|\frac{P(\sum_{j=1}^{k+1}\hat{X}_j>b_{1,k+1})}{P(\sum_{j=1}^{k+1}\tilde{X}_j>b_{1,k+1})}-1\right|\to 0$$

Thus, we can get (1). In addition, based on the similar arguments, we can get (2).

■

Proof [Proof of Lemma 8] (1) Let $Z_1',...,Z_K' \overset{iid}{\sim} N(0,1)$, with $2\leq K<M^{L-1}$. Define the set $\mathfrak{M}=\{\mathcal{M}_1\subset\{1,...,m\}: 1\leq|\mathcal{M}_1|\leq K-1\}$. It is suffice to show:

$$\lim_{m\to\infty}\sup_{\mathcal{M}_1\in\mathfrak{M}}\sup_{\substack{c_1\in[\beta_0,\gamma_m]\\ c_2\in[0,\gamma_m]}}\frac{\mathrm{P}(\frac{1}{\sqrt{K}}\sum_{i=1}^K Z_i'>c_2,\frac{1}{\sqrt{|\mathcal{M}_1|}}\sum_{j\in\mathcal{M}_1}Z_j'>c_1)}{\mathrm{P}(\frac{1}{\sqrt{K}}\sum_{i=1}^K Z_i'>c_2)}=0$$

Here, $\beta_0=\sqrt{2b(1-r_1)\log m+b(1-r_1)\log\log\log m}$, with

$$b=\frac{\frac{2M^{L-1}+1}{M^{L-1}+1}-r_1}{2(1-r_1)}\in\left(\frac{M^{L-1}}{(M^{L-1}+1)(1-r_1)},1\right).$$

For simplification, let $k_1=|\mathcal{M}_1|$. For Z_1 and $Z_2\overset{iid}{\sim}\mathrm{N}(0,1)$, define

$$\mathcal{D}_m=\left\{c_2\in(0,\gamma_m): \frac{d}{dc_2}\frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2, Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2\big)}=0\right\}$$

, then

$$\begin{aligned}
&\sup_{\substack{c_1\in[\beta_0,\gamma_m]\\ c_2\in[0,\gamma_m]}} \frac{P(\frac{1}{\sqrt{K}}\sum_{i=1}^{K} Z_i' > c_2, \frac{1}{\sqrt{|\mathcal{M}_1|}}\sum_{j\in\mathcal{M}_1} Z_j' > c_1)}{P(\frac{1}{\sqrt{K}}\sum_{i=1}^{K} Z_i' > c_2)}\\
\leq& 2\sup_{c_2\in[0,\gamma_m]} \frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)}\\
\leq& 2\max\Bigg\{\max_{c_2=0 \text{ or } \gamma_m} \frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)},\\
&\sup_{c_2\in\mathcal{D}_m} \frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)}\Bigg\}
\end{aligned}$$

(i). When $c_2 = 0$,

$$\lim_{m\to\infty} \frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)} = \lim_{m\to\infty} 2P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big) = 0$$

(ii). When $c_2 = \gamma_m$, $c_2/\beta_0 = \sqrt{\frac{1}{b(1-r_1)}}$,

$$\begin{aligned}
&\lim_{m\to\infty} \frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)} = \lim_{\beta_0\to\infty} \frac{\int_{\beta_0}^{\infty}\int_{S\sqrt{\frac{K}{K-k_1}}\beta_0-\sqrt{\frac{k_1}{K-k_1}}z_1}^{\infty}\phi(z_1)\phi(z_2)dz_2dz_1}{\int_{S\beta_0}^{\infty}\phi(z)dz}\\
\leq& C\lim_{\beta_0\to\infty} \frac{\int_{S\sqrt{\frac{K}{K-k_1}}\beta_0-\sqrt{\frac{k_1}{K-k_1}}\beta_0}^{\infty}\phi(\beta_0)\phi(z)dz + \int_{\beta_0}^{\infty}\phi(z)\phi(S\sqrt{\frac{K}{K-k_1}}\beta_0-\sqrt{\frac{k_1}{K-k_1}}z)dz}{\phi(S\beta_0)} \text{ (L'Hopital's rule)}\\
\leq& C\lim_{\beta_0\to\infty}\Bigg[\exp\Big\{-\frac{\beta_0^2}{2}\Big(S\sqrt{\frac{k_1}{K-k_1}}-\sqrt{\frac{K}{K-k_1}}\Big)^2\Big\} + \int_{\beta_0}^{\infty}\exp\Big\{-\frac{1}{2}\Big(\sqrt{\frac{K}{K-k_1}}z - S\sqrt{\frac{k_1}{K-k_1}}\beta_0\Big)^2\Big\}dz\Bigg] = 0,
\end{aligned}$$

Where $S = \sqrt{\frac{1}{b(1-r_1)}}$

(iii). When $c_2 \in \mathcal{D}_m$, given

$$\begin{aligned}
0 =& \frac{d}{dc_2}\frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)}\\
=& \frac{1}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)^2}\times\\
&\Big\{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big)\\
&- P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1>\beta_0\big)\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)\Big\}
\end{aligned}$$

We have

$$\frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2\big)}=\frac{\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0\big)}{\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2\big)}$$

Therefore,

$$\begin{aligned}
&\sup_{c_2\in\mathcal{D}_m}\frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2\big)}\\
=&\sup_{c_2\in\mathcal{D}_m}\frac{\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0\big)}{\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2\big)}\\
=&\sup_{c_2\in\mathcal{D}_m}C\int_{\beta_0}^{\infty}\exp\Big\{-\frac{1}{2}\Big(\sqrt{\frac{K}{K-k_1}}z-\sqrt{\frac{k_1}{K-k_1}}c_2\Big)^2\Big\}dz\\
\le&C\int_{\beta_0}^{\infty}\exp\Big\{-\frac{1}{2}\Big(\sqrt{\frac{K}{K-k_1}}z-\sqrt{\frac{k_1}{K-k_1}}\gamma_m\Big)^2\Big\}dz\\
\to&0
\end{aligned}$$

Combine (i), (ii) and (iii), we have

$$\lim_{m\to\infty}\sup_{\mathcal{M}_1\in\mathfrak{M}}\sup_{\substack{c_1\in[\beta_0,\gamma_m]\\c_2\in[0,\gamma_m]}}\frac{P(\frac{1}{\sqrt{K}}\sum_{i=1}^{K}Z_i>c_2|\frac{1}{\sqrt{|\mathcal{M}_1|}}\sum_{j\in\mathcal{M}_1}Z_j|>c_1)}{P(\frac{1}{\sqrt{K}}\sum_{i=1}^{K}Z_i>c_2)}=0$$

(2)

It is suffice to show

$$\lim_{m\to\infty}\sup_{\mathcal{M}_1\in\mathfrak{M}}\sup_{c_2\in[0,\bar{\Phi}^{-1}(1/m)]}\frac{\mathrm{P}(\frac{1}{\sqrt{K}}\sum_{i=1}^{K}X_i>c_2,\frac{1}{\sqrt{|\mathcal{M}_1|}}\sum_{j\in\mathcal{M}_1}X_j>\beta_0)}{\mathrm{P}(\sum_{i=1}^{K}Z_i'/\sqrt{K}>c_2)}\le 0$$

Let $\breve{X}_1=\sum_{i\in\mathcal{M}_1}X_i/\sqrt{k_1}$, $\breve{X}_2=\sum_{i\in\mathfrak{M}\setminus\mathcal{M}_1}X_i/\sqrt{K-k_1}$.

Based on lemma 7, $\delta_{6m}=|P(\breve{X}_j>p)/P(Z_j>p)-1|\to 0$ uniformly for $j=1,2$ and $p>\alpha_m$.

Thus, uniformly,

$$
\begin{aligned}
&P(\sqrt{\frac{k_1}{K}}\breve{X}_1+\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2,\breve{X}_1>\beta_0)\\
=&P(\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2-\sqrt{\frac{k_1}{K}}\beta_0,\breve{X}_1>\beta_0)\\
&+P(\sqrt{\frac{K-k_1}{K}}\breve{X}_2<c_2-\sqrt{\frac{k_1}{K}}\beta_0,\sqrt{\frac{k_1}{K}}\breve{X}_1+\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2)\\
\le&(1+\delta_{6m})\big[P(\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2-\sqrt{\frac{k_1}{K}}\beta_0,Z_1>\beta_0)\\
&+P(\sqrt{\frac{K-k_1}{K}}\breve{X}_2<c_2-\sqrt{\frac{k_1}{K}}\beta_0,\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2)+P(Z_1>\bar{\Phi}^{-1}(\alpha_m))\big]\\
\le&(1+\delta_{6m})^2\big[P(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0)\big]+(1+\delta_{6m})\sum_{j=1}^{2}P(Z_j>\bar{\Phi}^{-1}(\alpha_m))\\
\le&(1+\delta_{6m})^2\big[P(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0)\big]+2(1+\delta_{6m})\alpha_m\\
\le&o(\mathrm{P}(\sum_{i=1}^{K}Z_i'/\sqrt{K}>c_2))
\end{aligned}
$$

■

Proof [Proof of Lemma 9] **(i) Prove that (1) can leads to (2):**
On $\cap_{t=1}^{\ell}\mathcal{X}^{(t)}$,

$$\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I(T_S<\hat{t}^{(\ell)})\le\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}+\Big\{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}\Big\}\epsilon$$

Combined with

$$\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}\le\alpha\sum_{S\in\mathcal{B}^{(\ell)}}|S|\mathbb{I}\{T_S<\hat{t}^{(\ell)}\}$$

and

$$
\begin{aligned}
\sum_{S\in\mathcal{B}^{(\ell)}}|S|\mathbb{I}\{T_S<\hat{t}^{(\ell)}\}&=\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\mathbb{I}\{T_S<\hat{t}^{(\ell)}\}+\sum_{S\in\mathcal{B}_1^{(\ell)}}|S|\mathbb{I}\{T_S^{(\ell)}\le\hat{t}^{(\ell)}\}\\
&\le\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\mathbb{I}\{T_S^{(\ell)}<\hat{t}^{(\ell)}\}+Cm^{r_1}
\end{aligned}
$$

We have:

$$(1-\alpha-\alpha\epsilon)\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}\le\alpha Cm^{r_1}$$

Thus, $2|\mathcal{B}_0^{(\ell)}|\hat{t}^{(\ell)}\le\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}\le\frac{\alpha}{1-\alpha-\alpha\epsilon}m^{r_1}$, for any $1\le\ell\le L$.

When $\ell = 1$, by $|\mathcal{B}_0^{(1)}| = m_0 = m(1 + o(1))$, we have $\hat{t}^{(\ell)} \leq Cm^{(r_1-1)}$.

When $\ell \geq 2$, on $\cap_{k=1}^{(\ell)} \mathcal{X}^{(k)}$, we have

$$\max_{k=1,\ldots,\ell} \{FDP^{(k)} - \alpha\} < \epsilon$$

which leads to $|\mathcal{B}_0^{(\ell)}|/|\mathcal{B}^{(\ell)}| \to 1$. And accordingly, $\hat{t}^{(\ell)} \leq Cm^{(r_1-1)}$.

(ii) Prove that statement (2) leads to statement (3)

On layer 1, $\bar{\Phi}(\hat{c}_S) = \hat{t}^{(1)} \leq C(m)^{r_1-1}$. On layer $\ell \geq 2$ and $\cap_{h=1}^{\ell} \mathcal{X}^{(h)}$, for all $S \in \mathcal{B}^{(\ell)}$,

$$\bar{\Phi}(\hat{c}_S) \leq G_S(\hat{c}_S) + \sum_{S' \in \mathcal{U}(S)} \bar{\Phi}(\hat{c}_{S'}) \tag{S6}$$

Suppose $\bar{\Phi}(\hat{c}_{S'}) \leq C(m)^{r_1-1}$ for $S' \in \cup_{k=1}^{\ell-1} \mathcal{B}^{(k)}$, then together with $G_S(\hat{c}_S) = \hat{t}^{(\ell)} \leq Cm^{r_1-1}$ and (S6), we have

$$\hat{c}_S \geq \sqrt{2(1-r_1)\log m - 2\log\log m} = \beta_m$$

for all $S \in \mathcal{B}^{(\ell)}$.

In addition, for $S \in \mathcal{B}^{(\ell)}$, on $\cap_{h=1}^{\ell-1} \mathcal{X}^{(h)}$,

$$G_S(\hat{c}_S)[1 - \bar{\Phi}(\frac{\beta_0}{\sqrt{M^{L-1}}})]^{M^{L-1}} \leq \bar{\Phi}(\hat{c}_S) \tag{S7}$$

So we have $\bar{\Phi}(\hat{c}_S) \geq \hat{t}^{(\ell)}(1 + o(1))$, and accordingly, $\hat{c}_S \leq \gamma_m$.

Note that the $\hat{c}_S \leq \gamma_m$ only depends on the statement (2) on layer $\ell - 1$. Thus, we can apply the conclusion to show $\mathrm{P}(m_0\hat{t}^{(\ell)} > c\log m) \to 1$ in the proof of theorem 1.

(iii) Prove that statement (1) holds on layer 1 ($\ell = 1$):

Define $\nu_m = [(|\mathcal{A}'|^2/m + \delta_{2m}) \vee 1]/\sqrt{c_{\mathrm{md}} \log m}$. Let $0 = c_0 < ... < c_{\lceil \gamma_m/\nu_m \rceil} = \gamma_m$ satisfy $c_k - c_{k-1} = \nu_m$ for $1 \leq k < \lceil \gamma_m/\nu_m \rceil$ and $c_{\lceil \gamma_m/\nu_m \rceil} - c_{\lceil \gamma_m/\nu_m \rceil - 1} \leq \nu_m$. We can get the corresponding p-values sequence $q_0 > ... > q_{\lceil \gamma_m/\nu_m \rceil}$ with $q_k = 1 - \Phi(c_k)$. Let value $q^{(1)} = C^{(1)} c_{\mathrm{md}}/m$, by (S4), we have $P(\hat{t} > q^{(1)}) \to 1$. We define the working p-value sequence on layer 1 as $P_{sub}^{(1)} = \{q_0, ..., q_{k^{(1)}}, q^{(1)}\}$, where $k^{(1)} \in \{0, ..., \lceil \gamma_m/\nu_m \rceil - 1\}$ is the index s.t. $q_{k^{(1)}} \geq q^{(1)}$ and $q_{k^{(1)}+1} \leq q^{(1)}$.

If $\forall \epsilon > 0$,

$$P\Big(\max_{q \in P_{sub}^{(1)}} \Big| \frac{\sum_{S \in \mathcal{B}_0^{(1)}} I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S \in \mathcal{B}_0^{(1)}} P(X_S > \bar{\Phi}^{-1}(q))(1 - \delta_{0m})}{\sum_{S \in \mathcal{B}_0^{(1)}} q} \Big| > \epsilon \Big) \to 0 \tag{S8}$$

Then,

$$
\begin{aligned}
&P\bigg(\max_{q\in P_{sub}^{(1)}}\frac{\sum_{S\in\mathcal{B}_0^{(1)}} I(T_S^{(1)}<q)-\sum_{S\in\mathcal{B}_0^{(1)}} q}{\sum_{S\in\mathcal{B}_0^{(1)}} q}>\epsilon\bigg)\\
\le& P\bigg(\max_{q\in P_{sub}^{(1)}}\frac{\sum_{S\in\mathcal{B}_0^{(1)}} I(X_S>\bar\Phi^{-1}(q))-\sum_{S\in\mathcal{B}_0^{(1)}} P(\tilde X_S>\bar\Phi^{-1}(q))}{\sum_{S\in\mathcal{B}_0^{(1)}} q}>\epsilon\bigg)\\
\le& P\bigg(\max_{q\in P_{sub}^{(1)}}\frac{\sum_{S\in\mathcal{B}_0^{(1)}} I(X_S>\bar\Phi^{-1}(q))-\sum_{S\in\mathcal{B}_0^{(1)}} P(X_S>\bar\Phi^{-1}(q))(1-\delta_{0m})}{\sum_{S\in\mathcal{B}_0^{(1)}} q}>\epsilon\bigg)\\
\le& P\bigg(\max_{q\in P_{sub}^{(1)}}\bigg|\frac{\sum_{S\in\mathcal{B}_0^{(1)}} I(X_S>\bar\Phi^{-1}(q))-\sum_{S\in\mathcal{B}_0^{(1)}} P(X_S>\bar\Phi^{-1}(q))(1-\delta_{0m})}{\sum_{S\in\mathcal{B}_0^{(1)}} q}\bigg|>\epsilon\bigg)\\
=& o(1)
\end{aligned}
\tag{S9}
$$

Together with the fact that $\sup_{j=1,\dots,k}\left|q_{(j)}/q_{(j-1)}-1\right|=o(1)$, we have

$$
P\bigg(\sup_{q\in[q^{(1)},\alpha]}\frac{\sum_{S\in\mathcal{B}_0^{(1)}} I(T_S<q)-\sum_{S\in\mathcal{B}_0^{(1)}} q}{\sum_{S\in\mathcal{B}_0^{(1)}} q}>\epsilon\bigg)=o(1)
$$

Thus, to prove (1) holds on layer 1, we only need to show (S8).

Define $C_{sub}^{(1)}=\{c_0,...,c_{k'},c'\}$, with $c'=\bar\Phi^{-1}(q')$. In order to show (S8), it is suffice to show

$$
\int_0^{c'} P\bigg\{\bigg|\frac{\sum_{S\in\mathcal{B}_0^{(1)}} I(X_S>c)-P(X_S>c)(1-\delta_{0m})}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar\Phi(c)}\bigg|\ge\epsilon\bigg\}dc=o(\nu_m) \tag{S10}
$$

Note that by Markov inequality,

$$
\begin{aligned}
&P\bigg\{\bigg|\frac{\sum_{S\in\mathcal{B}_0^{(1)}}\big[I(X_S>c)-P(X_S>c)(1-\delta_{0m})\big]}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar\Phi(c)}\bigg|\ge\epsilon\bigg\}\\
\le& P\bigg\{\bigg|\frac{\sum_{S\in\mathcal{B}_0^{(1)}}\big[I(X_S>c)-P(X_S>c)\big]}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar\Phi(c)}\bigg|\ge\epsilon-(1+\delta_{0m})\delta_{0m}\bigg\}\\
\le&\frac{\sum_{S,S'\in\mathcal{B}_0^{(1)}}\big[P(X_S>c,X_{S'}>c)-P(X_S>c)P(X_{S'}>c)\big]}{\big(\sum_{S\in\mathcal{B}_0^{(1)}}\bar\Phi(c)\big)^2[\epsilon-(1+\delta_{0m})\delta_{0m}]^2}
\end{aligned}
$$

We can divide the $S,S'\in\mathcal{B}_0^{(1)}$ into the following three subsets:

$$
\begin{aligned}
\mathcal{B}_{01}^{(1)}&=\{S,S'\in\mathcal{B}_0^{(1)}:S=S'\}\\
\mathcal{B}_{02}^{(1)}&=\{S,S'\in\mathcal{B}_0^{(\ell)}:S\ne S',\exists A,A'\in\mathcal{A}^{(L)},s.t.S\subset A,S'\subset A',\text{ and }A'\in\Gamma_A\}\\
\mathcal{B}_{03}^{(1)}&=\{S,S'\in\mathcal{B}_0^{(1)}:S\ne S'\}\setminus\mathcal{B}_{02}^{(1)}
\end{aligned}
\tag{S11}
$$

Then,

$$\frac{\sum_{(S,S')\in\mathcal{B}_{01}^{(1)}}\left[P(X_S>c,X_{S'}>c)-P(X_S>c)P(X_{S'}>c)\right]}{\left(\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)\right)^2[\epsilon-(1+\delta_{0m})\delta_{0m}]^2}\leq\frac{C}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)}$$

Based on condition 3,

$$\frac{\sum_{(S,S')\in\mathcal{B}_{02}^{(1)}}\left[P(X_S>c,X_{S'}>c)-P(X_S>c)P(X_{S'}>c)\right]}{\left(\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)\right)^2[\epsilon-(1+\delta_{0m})\delta_{0m}]^2}\leq\frac{C(|\mathcal{A}'|^2/m+\delta_{2m})}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)}$$

In addition,

$$\frac{\sum_{(S,S')\in\mathcal{B}_{03}^{(1)}}\left[P(X_S>c,X_{S'}>c)-P(X_S>c)P(X_{S'}>c)\right]}{\left(\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)\right)^2[\epsilon-(1+\delta_{0m})\delta_{0m}]^2}=o(1)$$

Thus, after some calculation, we can prove (S10) and then $\mathrm{P}(\mathcal{X}^{(1)})\to 1$.

Similarly, if $|\tilde{\Omega}_0|=m(1+o(1))$, based on (S8), we have

$$P\left(\max_{q\in P_{sub}^{(1)}}\left|\frac{\sum_{S\in\mathcal{B}_0^{(1)}}I(T_S^{(1)}<q)-\sum_{S\in\mathcal{B}_0^{(1)}}q}{\sum_{S\in\mathcal{B}_0^{(1)}}q}\right|>\epsilon\right)=o(1)$$

Hence, $\mathrm{P}(\mathcal{X}'^{(1)})\to 1$.

(iv) Prove that statement (1) holds on layer $\ell\geq 2$ when statement (1) holds on previous layers:

On layer ℓ, we can divide the $S,S'\in\mathcal{B}_0^{(\ell)}$ into the following three subsets:

$$\begin{aligned}
\mathcal{B}_{01}^{(\ell)}&=\{S,S'\in\mathcal{B}_0^{(\ell)}:S=S',\{T_i:i\in S\}\text{ are mutually independent}\}\\
\mathcal{B}_{02}^{(\ell)}&=\{S,S'\in\mathcal{B}_0^{(\ell)}:\exists A,A'\in\mathcal{A}^{(L)},s.t.S\subset A,S'\subset A',\text{ and }A'\in\Gamma_A\}\\
\mathcal{B}_{03}^{(\ell)}&=\{S,S'\in\mathcal{B}_0^{(\ell)}:S\neq S'\}\setminus\mathcal{B}_{02}^{(\ell)}
\end{aligned}$$

Consider the p-values sequence $q_0>...>q_{\lceil\gamma_m/\nu_m\rceil}$ constructed in (iii). Let $q^{(\ell)}=C^{(\ell)}c_{\mathrm{md}}/m$, by (S4), we have $P(\hat{t}>q^{(\ell)})\to 1$. We define the working p-value sequence on layer 1 as $P_{sub}^{(\ell)}=\{q_0,...,q_{k^{(\ell)}},q^{(\ell)}\}$, where $k^{(\ell)}\in\{0,...,\lceil\gamma_m/\nu_m\rceil-1\}$ is the index s.t. $q_{k^{(\ell)}}\geq q^{(\ell)}$ and $q_{k^{(\ell)}+1}\leq q^{(\ell)}$.

In view of statement (3) and Lemma 8, we have

$$\sup_{k=0,...,\lceil\gamma_m/\nu_m\rceil}\left|\frac{G_S(c_k)}{\bar{\Phi}(c_k)}-1\right|=o(1)$$

Together with statement (3) and Lemma 7, there exists $\delta_5(m)\to 0$ with

$$\begin{aligned}
&\max_{S\in\mathcal{B}_0^{(\ell)}}\frac{\mathrm{P}(X_S>\bar{\Phi}^{-1}(q)|\mathcal{Q}^{(1:\ell-1)})}{q}\\
\leq&\max_{S\in\mathcal{B}_0^{(\ell)}}\frac{P(X_S>\bar{\Phi}^{-1}(q))}{P(Z_S>\bar{\Phi}^{-1}(q))[1-\bar{\Phi}(\frac{\beta_0}{\sqrt{M^{L-1}}})]^{M^{L-1}}}\\
\leq&1+\delta_5(m)
\end{aligned}$$

Then $\forall \epsilon > 0$, by following the similar arguments in (iii), we can have

$$P\Bigg(\max_{q \in P_{sub}^{(\ell)}} \left| \frac{\sum_{S \in \mathcal{B}_{01}^{(\ell)}} |S| I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S \in \mathcal{B}_{01}^{(\ell)}} |S| P(X_S > \bar{\Phi}^{-1}(q) | \mathcal{Q}^{(1:\ell-1)})(1+\delta_{0m})}{\sum_{S \in \mathcal{B}_{01}^{(\ell)}} |S| q} \right| > \epsilon \Bigg| \mathcal{Q}^{(1:\ell-1)} \Bigg)$$
$$\to 0 \tag{S12}$$

Then,

$$P\Bigg(\max_{q \in P_{sub}^{(\ell)}} \frac{\sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| I(T_S < q) - \sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| q}{\sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| q} > \epsilon \Bigg| \mathcal{Q}^{(1:\ell-1)} \Bigg)$$
$$\leq P\Bigg(\max_{q \in P_{sub}^{(\ell)}} \frac{\sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| P(X_S > \bar{\Phi}^{-1}(q) | \mathcal{Q}^{(1:\ell-1)})}{\sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| q} > \epsilon/2 \Bigg| \mathcal{Q}^{(1:\ell-1)} \Bigg)$$
$$= o(1) \tag{S13}$$

Together with the fact that $\sup_{j=1,\ldots,k} \left| q_{(j)}/q_{(j-1)} - 1 \right| = o(1)$, we have

$$P\Bigg(\sup_{q \in [q^{(\ell)}, \alpha]} \frac{\sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| I(T_S < q) - \sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| q}{\sum_{S \in \mathcal{B}_{0}^{(\ell)}} |S| q} > \epsilon \Bigg| \mathcal{Q}^{(1:\ell-1)} \Bigg) = o(1)$$

And thus $P(\mathcal{X}^{(\ell)}) \to 1$.

Similarly, when $|\tilde{\Omega}_0| = m(1+o(1))$, we have $P(\mathcal{X}'^{(\ell)}) \to 1$ based on Lemma 8 (2). ■

Proof [Proof of Lemma 10] When $\ell = 1$:

for $\delta = 1/m^4$,

$$\begin{aligned} \sum_{S \in \mathcal{B}_0^{(1)}} |S| \hat{t}^{(1)} &\leq \alpha \sum_{S \in \mathcal{B}^{(1)}} |S| I(T_S < \hat{t}^{(1)}) \\ &\leq \alpha \sum_{S \in \mathcal{B}^{(1)}} |S| I(T_S < \hat{t}^{(1)} + \delta) \\ &\leq \sum_{S \in \mathcal{B}_0^{(1)}} |S| \hat{t}^{(1)} (1 + o(1)) \end{aligned} \tag{S14}$$

Assume (12) holds on layer $1, \ldots, \ell - 1$. Then,

$$\sum_{S \in \mathcal{B}_0^{(\ell)}} |S| \hat{t}^{(\ell)} \leq \alpha(1+o(1)) \sum_{S \in \mathcal{B}^{(\ell)}} |S| I(T_S < \hat{t}^{(\ell)})$$

Thus, by following the similar arguments on (S14), we can get (12) on layer ℓ.

■