Improving online FDR procedures via online analogs of e-closure and compound e-values

Ziyu Xu[1] **Lasse Fischer**[2] **Aaditya Ramdas**[1]

[1]Carnegie Mellon University, USA
[2]University of Bremen, Germany

Abstract

In many scientific applications, hypotheses are generated and tested continuously in a stream. We develop a framework for improving online multiple testing procedures with false discovery rate (FDR) control under arbitrary dependence. Our approach is two-fold: we construct methods via the online e-closure principle, as well as a novel formulation of online compound e-values that is defined through donations. This yields strict power improvements over state-of-the-art e-value and p-value procedures while retaining FDR control. For standard immediate-decision online donation procedures, we further derive algorithms that compute the decision at time t in $O(\log t)$ time, and we demonstrate improved empirical performance on synthetic and real data.

1 INTRODUCTION

Large-scale hypothesis testing has become prevalent in multiple industries, such as A/B testing, genomics, clinical trials, neuroscience, and online monitoring, where the scientist wishes to filter for hypotheses where a true discovery is made and further investigation or action is merited. Online error control is particularly important in platform clinical trials (Robertson et al., 2023; Zehetmayer et al., 2022) and streaming anomaly detection (Lavin and Ahmad, 2015), where evidence arrives sequentially and testing decisions may trigger costly follow-up actions. Automated methods for doing so have become particularly prevalent as of late, with the rapidly improving capabilities of large language models (LLMs) to either act as an autonomous agent for generating hypotheses that it can then investigate, or as part of a human-in-the-loop system where the human scientist generates hypotheses and the model helps to triage them. In either case, hypotheses are often formulated in a sequential manner where the scientist (or LLM agent) generates some candidate hypotheses, gathers or analyzes data in the context of hypotheses, and then continues to generate more hypotheses to further elucidate their model of the world. Thus, one must provide statistical guardrails to such a system to ensure that the agent or system does not make too many false discoveries, and overfit their conclusions to the data at hand.

This motivates the problem of online multiple hypothesis testing (Foster and Stine, 2008), where we assume that there is an infinite stream of hypotheses $H_1, H_2, \dots$, and a new hypothesis arrives at each time step. As each hypothesis arrives, we assume we must make the decision whether to reject or accept the null hypothesis. Let $\mathcal{N} \subset \mathbb{N} := \{1, 2, \dots\}$ denote the set of null hypotheses, i.e., the subset of indices where the null hypothesis is true. Consequently, we output a monotonically growing sequence of discovery sets $\emptyset = R_0 \subseteq R_1 \subseteq R_2, \dots$ for each time step $t \in \mathbb{N}$, where R_t contains t if and only if we reject the tth null hypothesis.

Error metrics to control. The false positive criterion we wish to control is an online version of the false discovery rate (FDR) (Benjamini and Hochberg, 1995), an error criterion that has been central to statistical methodology in the offline multiple testing setting, where the number of hypotheses is known beforehand, for several decades. We define the FDR, along with the false discovery proportion (FDP), as follows:

$$\mathrm{FDP}_S(R) := \frac{|S \cap R|}{|R| \vee 1}, \qquad \mathrm{FDR}(R) := \mathbb{E}\left[\mathrm{FDP}_{\mathcal{N}}(R)\right].$$

In the above notation for FDP, S is a candidate set of null hypotheses and R is the discovery set the error metric is being evaluated on. The FDR is defined as the expectation of the FDP on the true null hypotheses $\mathcal{N}$. In the online setting, we output a novel discovery set at each time step t, which motivates the following online error metric that was recently proposed by Fischer et al. (2025):

$$\mathrm{SupFDR}(\mathbf{R}) := \mathbb{E}\left[\sup_{t \in \mathbb{N}} \mathrm{FDP}(R_t)\right].$$

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7552–7561.

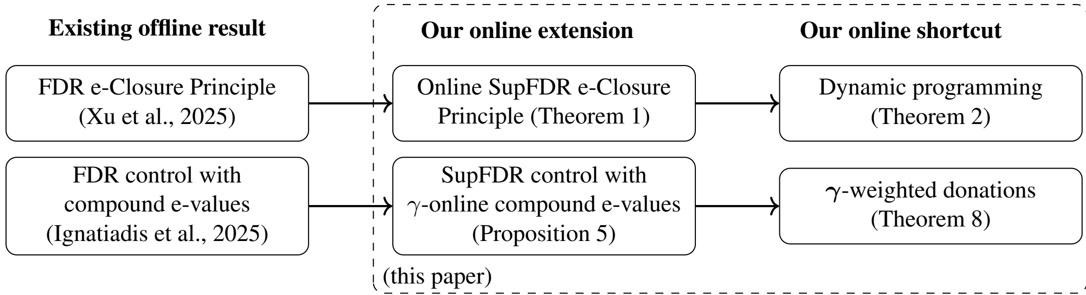


Figure 1: Summary of the paper's main technical contributions.

Online multiple testing with false discovery rate (FDR) control has been studied extensively in recent work (Xu and Ramdas, 2024). In this problem, we receive hypotheses in a stream, along with some data associated with those hypotheses. We then wish to produce a discovery set with control of the FDR at a fixed level $\delta \in (0, 1]$, and maximize the number of discoveries that we make. An interesting challenge that the online setting presents, in contrast to the classical offline multiple testing setting where one possesses a fixed, known number of hypotheses beforehand, is how to define the notion of FDR, now that one will output multiple discovery sets (usually one per time step).

Xu and Ramdas (2024) initiated a line of study in the use of e-values for onlineFDR control, a weaker form of FDR control for the online multiple testing setting (although Fischer et al. (2025) later showed that their methods controlled SupFDR as well), and assumed that the dependence between the data collected for different hypotheses was unknown. E-values are nonnegative random variables with expectation at most 1 under the null hypothesis, and have been shown to be useful for multiple testing in both offline and online settings — see Ramdas and Wang (2025) for an overview. Practically, many settings where online multiple testing is utilized (e.g., platform clinical trials (Robertson et al., 2023; Zehetmayer et al., 2022)) involve adaptive and sequential data collection, where e-values are more natural to construct than p-values (Ramdas et al., 2022). Further, offline e-closure ideas (Xu et al., 2025) suggest a unified way to characterize and improve e-value procedures, which we now adapt to the online SupFDR setting.

A brief overview of multiple testing. Multiple testing methods provide error guarantees when many hypotheses are tested together, and classically it is assumed that one knows the finite number of hypotheses a priori, i.e., offline multiple testing. The benchmark method for this setting is the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995), which is valid under independence or positive dependence among the p-values for each hypothesis, while the Benjamini-Yekutieli (BY) correction (Benjamini and Yekutieli, 2001) provides an analog that is valid under arbitrary dependence among the p-values. Another natural criterion for multiple testing is family-wise error rate (FWER) control, which controls the probability of making even one false discovery. FDR methods became popular as a less conservative alternative to FWER methods, since FWER control can be overly stringent in many applications (Benjamini and Hochberg, 1995; Lehmann and Romano, 2005). In online testing, hypotheses arrive sequentially and decisions must be made as the stream unfolds. In this setting, e-LOND is the standard e-value LOND procedure under arbitrary dependence, and r-LOND is a reshaped p-value LOND procedure that achieves arbitrary dependence validity by using reshaping functions. We recall both procedures formally below before presenting our improvements.

Contributions In this paper we introduce a framework that improves the power of existing online multiple testing procedures with FDR control. We first introduce an online e-closure principle for SupFDR control and apply it to improve a wide variety of online e-value and calibrated p-value procedures. This is not a direct reuse of the offline e-closure construction: the online setting requires an increasing e-collection over an infinite stream, validity uniformly over time, and test levels that can be recomputed sequentially using only past data. This methodology yields strict improvements over the status quo, but the resulting closed procedures are computationally expensive and may require $O(t^2)$ time to compute the rejection decision at time t. This quickly becomes costly in long streams.

Thus, another of the key contributions of this paper is deriving a practical algorithm that can improve power while being computationally tractable in the standard immediate-decision online setting. We provide a novel framework for generally improving e-value based online multiple testing procedures with FDR control, and show that it has practical power improvements over existing methods while requiring only $O(\log t)$ time to compute the rejection decision at time t for the donation e-LOND and donation r-LOND procedures. Our method is based on a novel notion of *online compound e-values*, which generalizes the notion of compound e-values for offline multiple testing (Ignatiadis et al., 2025) to the online setting. We show that online compound e-values can be used to generally improve online multiple testing procedures with FDR control, and demonstrate its empirical performance on both synthetic and real data. We also provide a user-friendly implementation of our methods in the accompanying code, and demonstrate that these donation methods are also computationally more efficient in practice.

Our primary contributions are as follows, and are also visually summarized in Figure 1.

1. *Closure-based strict power improvements for standard online testing.* We introduce an online e-closure principle for SupFDR control and use it to formulate strict improvements over e-LOND (Xu and Ramdas, 2024) and r-LOND (Zrnic et al., 2021). We derive novel explicit formulations for the next test level, which involves an optimization problem over subsets of the first $t-1$ hypotheses. Since naive optimization would require computation that is exponential in t, we also provide dynamic programming decomposition that only requires $O(t^2)$ time to compute the next test level.
2. *A computationally efficient donation framework for strict improvement.* Since $O(t^2)$ computation remains inefficient for moderately large values of t (i.e., in the thousands), we introduce a novel framework based on donations and online compound e-values, and show that it can also strictly improve e-LOND and r-LOND while remaining computationally tractable, requiring only $O(\log t)$ computation per time step. A central technical ingredient is our notion of *online compound e-values*, which powers the donation framework.
3. *Extensions beyond the standard online multiple testing setting.* Our donation framework is not restricted to only improving standard online multiple testing algorithms. We extend our framework to variants of the online multiple testing problem such as the acceptance-to-rejection setting of Fischer et al. (2025) and the decision deadlines setting of Fisher (2022), where we also derive strict improvements of existing algorithms. Lastly, we show that we can construct an efficient version of eBH (Wang and Ramdas, 2022) that is strictly more powerful for offline multiple testing, and is computationally more efficient than the $\overline{\text{eBH}}$ procedures of Xu et al. (2025); for m hypotheses, we require $O(m \log m)$ computation as opposed to the $O(m^2 \log m)$ required by $\overline{\text{eBH}}$.

We then demonstrate the empirical performance of our method on both synthetic and real data, where we see that both the power and computational improvements of the methods in this paper are nontrivial. Table 1 gives a compact summary of all improved procedures and their computational costs. By default, theorem proofs are deferred to Appendix B.

2 IMPROVING E-VALUE BASED PROCEDURES

Each arriving hypothesis H_t is paired with either an e-value E_t or a p-value P_t, depending on the procedure under consideration. Formally, the tth e-value and p-value satisfy the following properties, respectively:

$$\mathbb{E}[E_t] \leq 1 \text{ if } t \in \mathcal{N}. \quad (1)$$
$$\mathbb{P}(P_t \leq s) \leq s \text{ for all } s \in [0,1] \text{ if } t \in \mathcal{N}.$$

Setting	Base	Type	Method	Runtime
Standard	e-LOND	Closed	$\overline{\text{e-LOND}}$	$O(t^2)$
		Donation	donation e-LOND	$O(\log t)$
	r-LOND	Closed	$\overline{\text{r-LOND}}$	$O(t^2)$
		Donation	donation r-LOND	$O(\log t)$
ARC	online e-BH	Donation	donation online e-BH	$O(t \log t)$
Decision deadlines	e-TOAD	Donation	donation e-TOAD	$O(t \log t)$
Offline	e-BH	Closed	$\overline{\text{eBH}}$	$O(m^2 \log m)$
		Donation	donation e-BH	$O(m \log m)$

Table 1: Method summary of improved procedures grouped by setting and base procedure. For online methods, the “Runtime” column gives the asymptotic complexity of the tth online decision; cumulative costs over m online hypotheses are obtained by summing these per-decision costs. For offline methods, m is the number of hypotheses in the batch.

Online multiple testing algorithms can be viewed as producing a sequence of test levels $\boldsymbol{\alpha}$, where α_t is the level used at time t. For e-value procedures, we reject H_t when $E_t \geq \alpha_t^{-1}$; for p-value procedures, we reject H_t when $P_t \leq \alpha_t$. Thus, the induced discovery sets are:

$$R_t := \begin{cases} \{i \in [t] : E_i \geq \alpha_i^{-1}\} & \text{for e-values} \\ \{i \in [t] : P_i \leq \alpha_i\} & \text{for p-values} \end{cases},$$

where $[t] := \{1, \dots, t\}$. We say that a procedure with discovery sets $\mathbf{R}$ *strictly improves* another procedure with discovery sets $\mathbf{R}'$ if $R_t \supseteq R'_t$ for all $t \in \mathbb{N}$ almost surely, and there exists a data-generating distribution such that $\mathbb{P}(\exists t \in \mathbb{N} : R_t \supset R'_t) > 0$. Under unknown dependence between e-values, e-LOND controls SupFDR (Xu and Ramdas, 2024; Fischer et al., 2025). The e-LOND procedure uses a sequence of test levels $\boldsymbol{\alpha}$ defined by

$$\alpha_t := \delta\gamma_t(|R_{t-1}|+1), \quad (2)$$

where $\boldsymbol{\gamma}$ is a fixed nonnegative sequence of user-chosen constants such that $\sum_{t \in \mathbb{N}} \gamma_t \leq 1$. Similarly, Zrnic et al. (2021); Javanmard and Montanari (2018) showed that the following sequence of test levels $\boldsymbol{\alpha}$ ensures FDR control for p-values with arbitrary dependence:

$$\alpha_t := \delta\gamma_t\beta_t(|R_{t-1}| + 1), \quad (3)$$

where (β_t) is a sequence of reshaping functions (Blanchard and Roquain, 2008). A function $\beta : [0, \infty) \to [0, \infty)$ is a *reshaping function* if β can be written as $\beta(r) = \int_0^r x \, d\nu(x)$ where ν is any probability measure on $[0, \infty)$. A typical choice, which is the online analog of the Benjamini and Yekutieli (2001) correction for offline multiple testing under arbitrary dependence, is $\beta_t(r) = (\lfloor r \rfloor \wedge t)/\ell_t$, where $\ell_t := \sum_{i \in [t]} i^{-1}$ is the tth harmonic number.

We will demonstrate how to improve on both of these methods in the online setting in this paper, among others.

2.1 THE ONLINE SUPFDR E-CLOSURE PRINCIPLE

We develop an online e-closure principle that improves on existing methods in the online setting. Recently, Xu et al. (2025) proposed an e-Closure Principle and used this to improve the e-BH and the BY procedure for offline FDR control. Fischer et al. (2024) generalized the classical Closure Principle for FWER control (Marcus et al., 1976) to the online setting by using increasing families of local tests. In the following, we extend these ideas to introduce a SupFDR e-Closure Principle for the online setting.

Let $(\mathcal{F}_t)_{t\in\mathbb{N}}$ be a filtration where $\mathcal{F}_t$ denotes information available by time t. Let $\sigma(X)$ denote the sigma-algebra formed by a set X. Thus, we define each element of the filtration as the sigma-algebra $\mathcal{F}_t = \sigma(\{E_i\}_{i\in[t]})$ if one is working with e-values, and $\mathcal{F}_t = \sigma(\{P_i\}_{i\in[t]})$ if one is working with p-values. For the purposes of this framework, we consider a more general type of online procedure that outputs a collection $\mathcal{C}_t \subseteq 2^{[t]}$ of candidate rejection sets at each time t, with $\mathcal{C}_t$ being measurable w.r.t. $\mathcal{F}_t$. An e-collection is a family $(E_S)_{S\subset\mathbb{N},\,|S|<\infty}$, with $E_\emptyset = 0$, such that each E_S is an e-value for the intersection null $H_S = \cap_{i\in S} H_i$ when S is nonempty. In our online setting, we require *increasing* e-collections:

$$E_S \le E_{S\cup S'} \text{ for all } S, S' \subset \mathbb{N} \text{ s.t. } \min S' > \sup S.$$

Given an increasing e-collection, define:

$$\mathcal{C}_t := \left\{ R \subseteq [t] : E_S \ge \frac{\mathrm{FDP}_S(R)}{\delta} \text{ for all } S \subseteq [t] \right\}. \quad (4)$$

When (E_S) is increasing, the collections are nested ($\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \dots$), which is needed to obtain SupFDR control.

Theorem 1 (Online SupFDR e-closure). *Let $(E_S)_{S\subset\mathbb{N},\,|S|<\infty}$ be an increasing e-collection. Assume E_S is measurable with respect to $\mathcal{F}_{\sup(S)}$ for all finite nonempty S. Then the associated e-closure collections $(\mathcal{C}_t)_{t\in\mathbb{N}}$ in* (4) *form an online procedure that satisfies*

$$\mathbb{E}\left[\sup_{t\in\mathbb{N}} \sup_{R\in\mathcal{C}_t} \mathrm{FDP}_{\mathcal{N}}(R)\right] \le \delta.$$

Consequently, any discovery sequence $\mathbf{R}$ with $R_t \in \mathcal{C}_t$ for all $t \in \mathbb{N}$ controls SupFDR at level δ.

Proof details are deferred to Appendix B.1. When working with a stream of e-values, one explicit increasing e-collection is

$$E_S = \sum_{i\in S} \gamma_{|S\cap[i]|} E_i \text{ for all finite } S \subset \mathbb{N}. \quad (5)$$

For example, $E_{\{2,3\}} = \gamma_1 E_2 + \gamma_2 E_3$. This e-collection is increasing: if we append indices larger than $\max S$, the existing summands stay unchanged and we only add non-negative terms. As a result, we get the following constraint on E_t for making a discovery at time t, i.e., for $R_{t-1} \cup \{t\}$ to be in $\mathcal{C}_t$:

$$\mathrm{FDP}_S(R_{t-1} \cup \{t\}) \le \delta E_S \text{ for all } S \in 2^{[t]}.$$

With this closure framework in place, we now begin our first construction: applying the above increasing e-collection to derive a procedure that strictly improves over e-LOND. To see the resulting level, the constraints with $t \notin S$ are already implied by $R_{t-1} \in \mathcal{C}_{t-1}$. For the constraints with $t \in S$, write $S = S' \cup \{t\}$ with $S' \subseteq [t-1]$ and solve the inequality

$$\frac{1 + |S' \cap R_{t-1}|}{|R_{t-1}| + 1} \le \delta \left(E_{S'} + \gamma_{|S'|+1} E_t\right)$$

for E_t. Taking the largest required lower bound over S' yields the following test levels; we call the resulting procedure $\overline{\text{e-LOND}}$ *(closed e-LOND)*.

$$\alpha_t = \min_{\substack{S\subseteq[t-1]:\\ D_t(S)>0}} \frac{\delta\gamma_{|S|+1}(|R_{t-1}|+1)}{D_t(S)}, \quad (6)$$
$$D_t(S) := 1 + |S \cap R_{t-1}| - \delta E_S(|R_{t-1}| + 1).$$

Theorem 2. *The $\overline{\text{e-LOND}}$ procedure defined by the test levels in* (6) *ensures SupFDR control at level δ under arbitrary dependence between e-values, and dominates e-LOND when $\boldsymbol{\gamma}$ is nonincreasing. It strictly improves over e-LOND when the sequence has a strict decrease, i.e., instance $\gamma_i > \gamma_{i+!}$ for some $i \in \mathbb{N}$.*

Proof. The preceding derivation gives the exact threshold required for $R_{t-1} \cup \{t\}$ to satisfy all closure constraints involving the new index t. The constraints not involving t are already satisfied because $R_{t-1} \in \mathcal{C}_{t-1}$. Thus, whenever $\overline{\text{e-LOND}}$ rejects at time t, we have $R_t = R_{t-1} \cup \{t\} \in \mathcal{C}_t$; if it does not reject, then $R_t = R_{t-1} \in \mathcal{C}_t$ by monotonicity of the closure constraints. By induction, $R_t \in \mathcal{C}_t$ for all t, so SupFDR control follows from Theorem 1.

It remains to compare the test levels with e-LOND. Since $R_{t-1} \in \mathcal{C}_{t-1}$, for every $S \subseteq [t-1]$ we have $\mathrm{FDP}_S(R_{t-1}) \le \delta E_S$, and hence

$$|S \cap R_{t-1}| - \delta E_S |R_{t-1}| \le 0.$$

For any nonvacuous constraint in (6), the denominator is therefore at most one, while the nonincreasingness of $\boldsymbol{\gamma}$ gives $\gamma_{|S|+1} \ge \gamma_t$. Each candidate level is consequently at least $\delta\gamma_t(|R_{t-1}| + 1)$, the e-LOND level in (2). Hence $\overline{\text{e-LOND}}$ dominates e-LOND.

Strict improvement is possible already at $t = 2$. On an event with positive probability where $E_1 = (\delta\gamma_1)^{-1}$, both methods reject H_1, and the $S = \{1\}$ constraint in (6) is vacuous because

$$1 + |S \cap R_1| - \delta E_S(|R_1| + 1) = 2 - 2\delta\gamma_1 E_1 = 0.$$

Thus $\alpha_2^{\overline{\text{e-LOND}}} = 2\delta\gamma_1 > 2\delta\gamma_2 = \alpha_2^{\text{e-LOND}}$ when $\gamma_1 > \gamma_2$. A two-point construction with positive probability on this value of E_1, and with E_2 falling between the two rejection thresholds with positive probability, gives an explicit distribution on which $\overline{\text{e-LOND}}$ rejects H_2 while e-LOND does not. □

Computation via dynamic programming. Although we have shown that $\overline{\text{e-LOND}}$ strictly improves over e-LOND, it is expensive to compute. While Xu et al. (2025) developed computational shortcuts for the offline e-closure principle, it is not obvious how to adapt such shortcuts to the online setting due to the weighted merging of e-values involved in solving the maximization problem in (6). A naive computation would require exponential time in t, while dynamic programming reduces this to a worst-case $O(t^2)$ computation per time step. To carry out this computation, we define:

$$v_t(i,k) := \max_{S\subseteq[i]:|S|=k} |S\cap R_{t-1}| - \delta E_S(|R_{t-1}|+1).$$

The $\overline{\text{e-LOND}}$ choice of α_t in (6) can be computed as

$$\alpha_t = \min_{\substack{k\in\{0\}\cup[t-1]:\\ 1+v_t(t-1,k)>0}} \frac{\delta\gamma_{k+1}(|R_{t-1}|+1)}{1+v_t(t-1,k)}.$$

We can compute $v_t(i,k)$ for $i\in[t-1]$ and $k\in\{0\}\cup[t-1]$ using the following dynamic programming formula (for $k\geq 1$, and with $v_t(i,0)=0$):

$$\begin{aligned} v_t(i,k) = \max\{&v_t(i-1,k),\\ &v_t(i-1,k-1) + \mathbf{1}\{i\in R_{t-1}\} - \delta\gamma_k E_i(|R_{t-1}|+1)\}. \end{aligned}$$

Thus, computing the dynamic program requires $O(t^2)$ time. However, this quickly becomes computationally costly in practice, motivating the need for more efficient methods.

Remark 3. One minor drawback of the e-collection in (5) is that showing strict improvement in Theorem 2 requires $\boldsymbol{\gamma}$ to be nonincreasing. To avoid this restriction, we can define

$$E_S = \sum_{i\in S}\gamma_{i,S}E_i,$$

where $\gamma_{i,S} := \sum_{j\in N_S(i)}\gamma_j$ and $N_S(i) := \{\max(S\cap[i-1])+1,\ldots,i\}$, with the convention $\max(\emptyset)=0$. Intuitively, this is still a weighted sum of the e-values in S, but each weight aggregates γ-mass between consecutive selected indices in time order. This requires different computational shortcuts for test-level computation; see Appendix A.2.

2.2 COMPOUND E-VALUES VIA DONATION

So far, online multiple testing methods have primarily treated e-values either as direct inputs or as intermediaries for improving p-value procedures. However, we will instead leverage the notion of *compound e-values* to substantially improve power. In offline multiple testing with a fixed $m\in\mathbb{N}$, one calls nonnegative random variables $(\tilde{E}_1,\ldots,\tilde{E}_m)$ *compound e-values* if $\sum_{i\in\mathcal{N}}\mathbb{E}[\tilde{E}_i]\leq m$ (Ignatiadis et al., 2025). This relaxes the usual e-value condition in (1). We now introduce the online, $\boldsymbol{\gamma}$-weighted analog together with donation sequences used to construct it. Fix a nonnegative sequence $\boldsymbol{\gamma}=(\gamma_t)_{t\in\mathbb{N}}$ with $\sum_{t\in\mathbb{N}}\gamma_t\leq 1$.

Definition 4 ($\boldsymbol{\gamma}$-online compound e-values and $\boldsymbol{\gamma}$-weighted donations)**.** For this fixed $\boldsymbol{\gamma}$:

(i) $\tilde{\mathbf{E}}=(\tilde{E}_t)_{t\in\mathbb{N}}$ is a stream of $\boldsymbol{\gamma}$*-online compound e-values* if $\sum_{i\in\mathcal{N}}\gamma_i\mathbb{E}[\tilde{E}_i]\leq 1$.

(ii) $\mathbf{B}=(B_t)_{t\in\mathbb{N}}$ is a $\boldsymbol{\gamma}$*-weighted donation* if $\sum_{i\in[t]}\gamma_iB_i\leq 0$ for all $t\in\mathbb{N}$, and $B_t\geq -(E_t\wedge 1)$ for all $t\in\mathbb{N}$.

We first note that weighted self-consistency applied to online compound e-values has valid SupFDR control. Define the following collection of finite weighted self-consistent discovery sets:

$$\mathcal{C} := \left\{R\subset\mathbb{N}: |R|<\infty,\ \tilde{E}_t\geq\frac{1}{\delta\gamma_t|R|}\text{ for all } t\in R\right\}, \tag{7}$$

where $(\delta\gamma_t|R|)^{-1}$ is interpreted as $+\infty$ when $\gamma_t=0$.

Proposition 5. *Let $\tilde{E}_1,\tilde{E}_2,\ldots$ be a stream of $\boldsymbol{\gamma}$-online compound e-values. Then,*

$$\mathbb{E}\left[\sup_{R\in\mathcal{C}}\text{FDP}(R)\right]\leq\delta.$$

A full proof is provided in Appendix B.2. Now that we have shown that online compound e-values can be used to control SupFDR, we introduce a construction that, when combined with the test levels of e-LOND, strictly dominates e-LOND applied only to the original e-values. This construction is computationally efficient, requiring only $O(\log t)$ time to compute the compound e-value and hence the rejection decision at each time step. Further, it is robust to unknown dependence between e-values. Now, we will define how to construct online compound e-values via *donation*. Let $\mathbf{B}$ be any $\boldsymbol{\gamma}$-weighted donation sequence as in Definition 4. Note that $\mathbf{B}$ can be arbitrarily dependent with $\mathbf{E}$. We first note the following property.

Proposition 6. *Let $\tilde{E}_t=E_t+B_t$. Then, for all $t\in\mathbb{N}$, $\tilde{\mathbf{E}}$ is a valid sequence of $\boldsymbol{\gamma}$-online compound e-values.*

Proof details are deferred to Appendix B.3.

As a result, we can choose any $\boldsymbol{\gamma}$-weighted donation $\mathbf{B}$ to construct online compound e-values $\tilde{\mathbf{E}}$ for e-LOND while retaining SupFDR control. Furthermore, we can take a supremum over choices of $\mathbf{B}$ and still retain FDP control. Let $\mathcal{C}(\mathbf{B})$ be the collection of finite discovery sets defined by the online weighted self-consistency condition in (7) with online compound e-values $\tilde{E}_t=E_t+B_t$ for a specific choice of $\mathbf{B}$. For a stream of arbitrarily dependent e-values $\mathbf{E}$, let $\mathcal{B}$ be the set of all valid $\boldsymbol{\gamma}$-weighted donations. Then we have:

Proposition 7. *For any stream of e-values $E_1, E_2, \ldots$, we have that*

$$\mathbb{E}\left[\sup_{\mathbf{B}\in\mathcal{B}} \sup_{R\in\mathcal{C}(\mathbf{B})} \mathrm{FDP}(R)\right] \le \delta.$$

See Appendix B.4 for the proof.

As a result, we can define an algorithm that is equivalent to choosing the following test levels. Define the following "wealth" quantity for each $t \in \mathbb{N}$:

$$\begin{aligned}\bar{W}_t &:= \sum_{i\in R_{t-1}} \gamma_i \left(\left(E_i - \frac{1}{\delta\gamma_i(|R_{t-1}|+1)}\right) \wedge 1\right) \\ &\quad + \sum_{i\in[t-1]\setminus R_{t-1}} \gamma_i(E_i \wedge 1),\end{aligned}$$

where $R_{t-1} = \{i \in [t-1] : E_i \ge \alpha_i^{-1}\}$ is defined by the following test levels:

$$\alpha_t := \frac{\delta\gamma_t(|R_{t-1}|+1)}{1-(\delta(|R_{t-1}|+1)\bar{W}_t \wedge 1)}. \tag{8}$$

We refer to this procedure as *donation e-LOND*. The convention $R_0 = \emptyset$ makes $\bar{W}_1 = 0$ and hence $\alpha_1 = \delta\gamma_1$. Intuitively, $\bar{W}_t$ is the largest γ-weighted amount of past evidence that can be shifted to E_t while keeping all previous discoveries in R_{t-1} valid and respecting the donation budget. Thus, previous e-values that exceed what is needed for their current rejection can donate excess mass, and unrejected e-values can donate up to $E_i \wedge 1$.

Note that $(1-(\delta(|R_{t-1}|+1)\bar{W}_t \wedge 1))^{-1} \ge 1$, so donation e-LOND always has test levels at least as large as e-LOND. The two improvements should be viewed as complementary: neither donation e-LOND nor $\overline{\text{e-LOND}}$ dominates the other pointwise, although both dominate e-LOND. We now have the following result.

Theorem 8. *Donation e-LOND controls the SupFDR at level δ, and strictly improves over e-LOND.*

Proof sketch. Suppose at time t that we want the enlarged rejection set $R_{t-1} \cup \{t\}$ to be weighted self-consistent for some valid $\boldsymbol{\gamma}$-weighted donation $\mathbf{B}$. This requires

$$E_t + B_t \ge \frac{1}{\delta\gamma_t(|R_{t-1}|+1)}, \tag{9}$$

$$E_i + B_i \ge \frac{1}{\delta\gamma_i(|R_{t-1}|+1)} \qquad \text{for each } i \in R_{t-1}. \tag{10}$$

The donation budget gives $\gamma_t B_t \le -\sum_{i<t}\gamma_i B_i$. To make rejection of H_t as easy as possible, we maximize the donation available to t by choosing the smallest feasible B_i for each past index. For $i \in R_{t-1}$, (10) and $B_i \ge -(E_i \wedge 1)$ imply

$$B_i^{\min} = \left(\frac{1}{\delta\gamma_i(|R_{t-1}|+1)} - E_i\right) \vee (-1),$$

so

$$-\gamma_i B_i^{\min} = \gamma_i \wedge \left(\gamma_i E_i - \frac{1}{\delta(|R_{t-1}|+1)}\right).$$

For $i \notin R_{t-1}$, there is no rejection-preserving constraint, so $B_i^{\min} = -(E_i \wedge 1)$ and $-\gamma_i B_i^{\min} = \gamma_i(E_i \wedge 1)$. Summing these maximal past contributions yields exactly $\bar{W}_t$, and hence the largest feasible donation to the new hypothesis is $\gamma_t B_t = \bar{W}_t$, up to the point where the rejection constraint is already vacuous. Substituting $B_t = \bar{W}_t/\gamma_t$ into (9) gives

$$\begin{aligned}E_t &\ge \frac{1}{\delta\gamma_t(|R_{t-1}|+1)} - \frac{\bar{W}_t}{\gamma_t} \\ &= \frac{1-\delta(|R_{t-1}|+1)\bar{W}_t}{\delta\gamma_t(|R_{t-1}|+1)}.\end{aligned}$$

When $\delta(|R_{t-1}|+1)\bar{W}_t \ge 1$, the right-hand side is nonpositive, so any nonnegative e-value satisfies the constraint. This is captured by replacing the numerator with $1-(\delta(|R_{t-1}|+1)\bar{W}_t \wedge 1)$, which is equivalent to the test level in (8). SupFDR control then follows from Proposition 7; the strict improvement over e-LOND follows from the displayed test level being at least the e-LOND level, with strict inequality possible when past evidence can donate positive mass. □

The proof appears in Appendix B.5.

Efficient donation computation. To efficiently compute the test levels in (8), we need to update $\bar{W}_t$ efficiently at each time step. The only nontrivial component to compute is the summation over terms in R_{t-1}, i.e., our term of interest is

$$\begin{aligned}\bar{W}_t^R &:= \sum_{i\in R_{t-1}} \gamma_i \left(\left(E_i - \frac{1}{\delta\gamma_i(|R_{t-1}|+1)}\right) \wedge 1\right) \\ &= \sum_{i\in R_{t-1}} \gamma_i \wedge \left(\gamma_i E_i - \frac{1}{\delta(|R_{t-1}|+1)}\right).\end{aligned}$$

Define

$$\begin{aligned}\bar{w}_t^{(i)} &:= \gamma_i \wedge \left(\gamma_i E_i - \frac{1}{\delta(|R_{t-1}|+1)}\right) \\ &= \begin{cases} \gamma_i E_i - \frac{1}{\delta(|R_{t-1}|+1)} & \text{if } \gamma_i(E_i-1) \le \frac{1}{\delta(|R_{t-1}|+1)} \\ \gamma_i & \text{otherwise.}\end{cases}\end{aligned}$$

As a result, we need to threshold on the value of $\gamma_i(E_i - 1)$ for each $i \in [t]$ to determine what value $\bar{w}_t^{(i)}$ takes on. To compute the sum of $\bar{w}_t^{(i)}$ efficiently, we maintain an augmented binary search tree where the key is $\gamma_i(E_i-1)$ for each $i \in R_{t-1}$. We augment each node with sums of $\gamma_i E_i$, γ_i and a count of nodes for all nodes i that are in the tree. Therefore, when split on $\gamma_i(E_i - 1)$, we have already computed our desired quantities. Consequently, the computation is simply $O(\log(|R_{t-1}|)) \le O(\log t)$ whenever we make a discovery, i.e., the insertion cost into the augmented binary search tree.

3 IMPROVING P-VALUE BASED PROCEDURES

Using the above results for improving e-value based procedures, we can also improve the r-LOND procedure for p-value based procedures. Xu and Ramdas (2024) observed that the r-LOND procedure was equivalent to applying e-LOND where the e-values were defined by $E_t = f_t(P_t)$, where we let f_t be the following calibrator similar to the calibrator for r-LOND Xu and Ramdas (2024) for each $t \in \mathbb{N}$:

$$f_t(p) = \frac{\mathbf{1}\{p \leq \delta\gamma_t t/\ell_t\}}{\delta\gamma_t \lceil (p\ell_t/(\delta\gamma_t)) \vee 1 \rceil}. \tag{11}$$

As a result, we can apply online e-closure principle to p-value based procedures and achieve SupFDR control. Notably, unlike the formulation of r-LOND as the application of e-LOND to calibrated p-values $f_t(P_t)$ for $t \in \mathbb{N}$, we instead construct a new e-collection. The calibrator index below is the rank $|S \cap [i]|$ of hypothesis i within the subset S, rather than its global time index i. This rank-based choice is what makes the dynamic program in Appendix A.3 depend only on the subset size, giving an $O(t^2)$ computation. Let

$$\begin{aligned} E_S &= \sum_{i \in S} \gamma_{|S\cap[i]|} f_{|S\cap[i]|}(P_i) \\ &= \frac{1}{\delta} \sum_{i \in S} \frac{\mathbf{1}\{P_i \leq \delta\gamma_{|S\cap[i]|} |S \cap [i]| / \ell_{|S\cap[i]|}\}}{\lceil (P_i \ell_{|S\cap[i]|}/(\delta\gamma_{|S\cap[i]|})) \vee 1 \rceil}. \end{aligned} \tag{12}$$

Consequently, we define $\overline{\textit{r-LOND}}$ *(closed r-LOND)* as

$$\alpha_t := 1 \wedge \min_{\substack{S \subseteq [t-1]: \\ \Delta E_S(t) > 0}} \frac{\delta\gamma_{|S|+1}}{\ell_{|S|+1}} \lfloor (\delta \Delta E_S(t))^{-1} \wedge (|S|+1) \rfloor \tag{13}$$

$$\text{where } \Delta E_S(t) := \frac{1 + |S \cap R_{t-1}|}{\delta(|R_{t-1}|+1)} - E_S.$$

Note that the constraint set is never empty since we can always select $S = \emptyset$.

Theorem 9. *$\overline{\text{r-LOND}}$ controls the SupFDR at level δ for arbitrarily dependent p-values. Further, when $i\gamma_i/\ell_i$ is non-increasing in i for $i \in \mathbb{N}$, $\overline{\text{r-LOND}}$ dominates r-LOND. It strictly improves over r-LOND when there is a strict decrease, i.e., $\gamma_i > \gamma_{i+1}$ for some $i \in \mathbb{N}$.*

The proof is deferred to Appendix B.6. We also elaborate on computational details for $\overline{\text{r-LOND}}$ in Appendix A.3, but they are similar to that of $\overline{\text{e-LOND}}$, and it consequently requires only $O(t^2)$ computation at each time step.

Define the donation excess-wealth term $\bar{W}_t$ as we do in (8), but using the calibrated e-values derived via $E_t = f_t(P_t)$. We then obtain the following test levels for arbitrarily dependent p-values:

$$\alpha_t := 1 \wedge \frac{\delta\gamma_t}{\ell_t} \left(\left\lfloor \frac{|R_{t-1}|+1}{1 - (\delta(|R_{t-1}|+1)\bar{W}_t \wedge 1)} \right\rfloor \wedge t \right). \tag{14}$$

Theorem 10. *Donation r-LOND controls the SupFDR at level δ for arbitrarily dependent p-values, and can strictly improve over r-LOND.*

A proof of the above is deferred to Appendix B.7.

4 SIMULATIONS

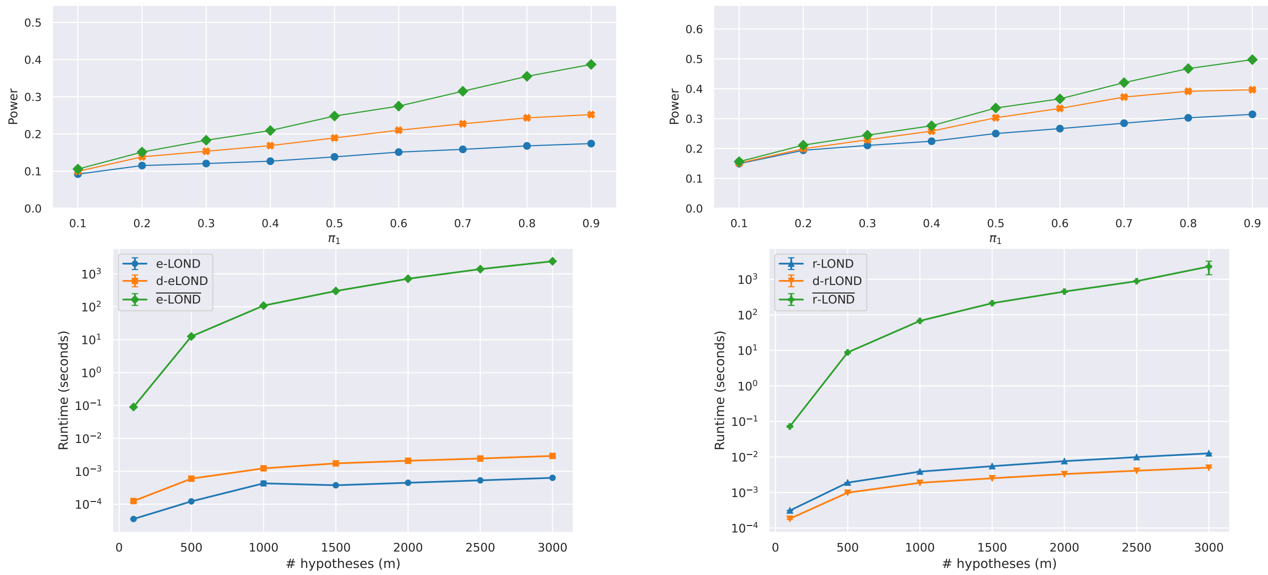
We compare our procedures in a local dependence setup inspired by Zrnic et al. (2021). Each hypothesis $t \in [m]$ produces a single Gaussian observation with $\mu_0 = 0$ under the null and $\mu_1 = 3$ under the alternative, where π_1 denotes the probability of the alternative being true; we set $m = 200$. For a fixed lag $L = 100$, samples within distance L of one another share Gaussian copula dependence: the covariance matrix satisfies $\Sigma_{i,j} = 0.5^{|i-j|}$ for $|i-j| \leq L$ and zero otherwise, and we verify positive semidefiniteness numerically for each simulated dimension. The resulting e-values and p-values are based on Gaussian likelihood ratios and the Gaussian c.d.f. — see Appendix C.3 for details. Each design is averaged over $n = 200$ trials. For all methods, we use the sequence where $\gamma_t = (t(t+1))^{-1}$ for each $t \in \mathbb{N}$ unless otherwise stated. We also consider additional simulation settings in Appendix C.1.

We see in Figure 2 the results of applying the base procedure of e-LOND and r-LOND, as well as the power gain our procedures offer over both e-LOND and r-LOND. Both donation e-LOND and $\overline{\text{e-LOND}}$ outperform e-LOND, with $\overline{\text{e-LOND}}$'s power differential increasing over donation e-LOND as the proportion of non-nulls increases. Similarly, both donation r-LOND and $\overline{\text{r-LOND}}$ also outperform r-LOND, with $\overline{\text{r-LOND}}$'s power differential increasing over donation r-LOND as the proportion of non-nulls increases. Thus, we see that our frameworks for improving the power of both procedures have practical results.

Runtime comparison To quantify the computation–power tradeoff, we run the same local dependence Gaussian simulation with $\pi_1 = 0.3$, hypothesis counts up to $m = 3000$, and $n = 80$ trials. Wall-clock times were measured in single-threaded runs on an AMD Ryzen 9 9950X CPU. Figure 2 shows that as m grows, $\overline{\text{e-LOND}}$ goes from milliseconds to roughly an hour, while e-LOND and donation e-LOND remain in the millisecond range; the r-LOND variants show the same pattern. In cumulative terms over m online decisions, closed variants scale as $O(m^3)$ when recomputed from scratch, while donation e-LOND and donation r-LOND scale as $O(m \log m)$ with the augmented-tree updates described above.

5 REAL DATA EXPERIMENT

We also evaluate our procedures on the NYC taxi demand anomaly-detection experiment used by Zhang et al. (2026). The NYC taxi dataset (Lavin and Ahmad, 2015) records taxi demand over time. We use the likelihood-ratio e-value setup of Zhang et al. (2026); details and annotation-window



(a) E-value procedures.

(b) P-value procedures.

Figure 2: Local dependence simulation summary over 200 trials. The left column shows e-value procedures and the right column shows p-value procedures. The top row reports power as the non-null fraction π_1 increases (with $\mu = 3$ and $\delta = 0.1$), while the bottom row reports mean wall-clock runtime (log scale) as the number of hypotheses increases. Donation and closed variants improve power over the corresponding baselines, and donation variants remain computationally practical compared with closed variants.

e-LOND	d-eLOND	$\overline{\text{e-LOND}}$
8	**16**	10

Table 2: Discoveries coinciding with annotated anomalous periods for each e-value procedure on the NYC taxi dataset. Donation e-LOND and $\overline{\text{e-LOND}}$ both improve over e-LOND, with donation e-LOND detecting the largest number of annotated anomalies.

interpretation are in Appendix C.2. Because the density ratio is fitted and normalized from the observed series, this is an empirical anomaly-detection illustration rather than a formal SupFDR validation under estimated density ratios. Table 2 reports rejections inside annotated anomaly windows. These annotations are incomplete event windows rather than exact true-discovery labels. Both $\overline{\text{e-LOND}}$ and donation e-LOND outperform e-LOND, with donation e-LOND having the largest gain. This also illustrates that $\overline{\text{e-LOND}}$ and donation e-LOND do not dominate each other, despite each strictly improving over e-LOND. We visualize the discoveries in Figure 3.

6 EXTENSIONS

The donation framework also improves three other classes of algorithms.

1. **Online acceptance-to-rejection (ARC) and decision deadlines.** Fischer et al. (2025) introduced the online ARC problem, where decisions need not be immediate. We use donations to efficiently improve online e-BH and the decision-deadline procedure of Fisher (2022); see Appendix D.
2. **Offline multiple testing.** Donations yield compound e-values for eBH, improving power while computing a strictly more powerful discovery set in $O(m \log m)$ time rather than the quadratic time of closed eBH; see Appendix D.3.
3. **Randomization**: Xu and Ramdas (2026) and Xu and Ramdas (2024) used randomization to improve multiple-testing procedures; we apply the same idea to donation procedures in Appendix E.

7 RELATED WORK

Online multiple testing with FDR control has been studied extensively. Javanmard and Montanari (2018) introduced online FDR control and LORD under independence, followed by increasingly powerful algorithms under independence or conditional validity assumptions (Ramdas et al., 2017; Zrnic et al., 2020). Ramdas et al. (2018) and Tian and Ramdas (2019) formulate adaptive and discarding LORD variants in the style of Storey (2002) and Zhao et al. (2019). These methods are powerful under their assumptions, but they are not valid in the arbitrary-dependence SupFDR setting considered here, so they are not direct baselines. More recently Zhang et al. (2026) proposed a conditional-validity method for p-values and e-values.

A line of work has also considered other dependence structures. In addition to defining r-LOND under arbitrary dependence, Zrnic et al. (2021) study global positive and local dependence; Fisher (2024) shows that LORD and SAFFRON are valid under another positive-dependence notion; and Jankovic et al. (2026) consider asymptotic online FWER

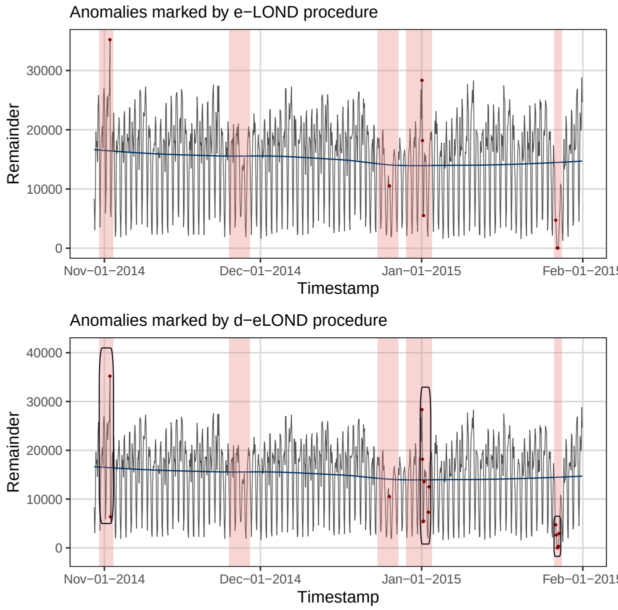


Figure 3: Plot of NYC taxi usage over time, with red highlighted periods indicating anomalous periods. We can see donation e-LOND making more discoveries in these periods than e-LOND (circled in black).

control for weakly dependent data.

Outside of Xu and Ramdas (2024), e-values have also been used for online FWER control: Fischer and Ramdas (2025) showed that admissible online methods controlling FDP tail probabilities must use e-values.

Different online multiple testing analogs of FDR. SupFDR is stronger than two classical online FDR criteria. The first is onlineFDR (Javanmard and Montanari, 2018):

$$\text{onlineFDR}(\mathbf{R}) := \sup_{t\in\mathbb{N}} \text{FDR}(R_t).$$

Fischer et al. (2025) also defined the stopping-time criterion

$$\text{StopFDR}(\mathbf{R}) := \sup_{\tau\in\mathcal{T}} \text{FDR}(R_\tau),$$

where $\mathcal{T}$ is the set of stopping times. SupFDR control implies both because $\sup_t \mathbb{E}[\text{FDP}(R_t)] \leq \mathbb{E}[\sup_t \text{FDP}(R_t)]$ and $\text{FDP}(R_\tau) \leq \sup_t \text{FDP}(R_t)$ for any stopping time τ.

8 CONCLUSION

We introduced a framework for improving online multiple testing procedures under unknown dependence, with a focus on SupFDR control. Online e-closure yields closed procedures that dominate their baselines, including e-LOND and r-LOND, but can be expensive in long streams. The donation framework instead constructs online compound e-values via γ-weighted donations, giving computationally efficient improvements such as donation e-LOND and donation r-LOND. Together, these approaches provide a principled tradeoff between power and computation. An important direction is to further narrow this computational gap, or show that it is irreducible.

References

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 (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.

Lasse Fischer and Aaditya Ramdas. Admissible online closed testing must employ e-values. arXiv:2407.15733, 2025.

Lasse Fischer, Marta Bofill Roig, and Werner Brannath. The online closure principle. *The Annals of Statistics*, 52(2): 817–841, 2024.

Lasse Fischer, Ziyu Xu, and Aaditya Ramdas. An online generalization of the (e-)Benjamini-Hochberg procedure. arXiv:2407.20683, 2025.

Aaron Fisher. Online false discovery rate control for LORD++ and SAFFRON under positive, local dependence. *Biometrical Journal*, 66(1):2300177, 2024.

Aaron J. Fisher. Online Control of the False Discovery Rate under "Decision Deadlines". In *International Conference on Artificial Intelligence and Statistics*, 2022.

Dean P. Foster and Robert A. Stine. α-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444, 2008.

Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. Asymptotic and compound e-values: Multiple testing and empirical Bayes. arXiv:2409.19812, 2025.

Vincent Jankovic, Lasse Fischer, and Werner Brannath. Asymptotic Online FWER Control for Dependent Test Statistics. *Statistical Methods in Medical Research (forthcoming)*, 2026.

Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, 46(2):526–554, 2018.

Alexander Lavin and Subutai Ahmad. Evaluating Real-Time Anomaly Detection Algorithms – The Numenta Anomaly Benchmark. In *IEEE International Conference on Machine Learning and Applications*, 2015.

E. L. Lehmann and Joseph P. Romano. Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3): 1138–1154, 2005.

Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2): 1–390, 2025.

Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. Online control of the false discovery rate with decaying memory. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aaditya Ramdas, Tijana Zrnic, Martin Wainwright, and Michael Jordan. SAFFRON: An Adaptive Algorithm for Online Control of the False Discovery Rate. In *International Conference on Machine Learning*, 2018.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv:2009.03167, 2022.

David S Robertson, James MS Wason, Franz König, Martin Posch, and Thomas Jaki. Online error rate control for platform trials. *Statistics in Medicine*, 42(14):2475–2495, 2023.

John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

Weijie J. Su. The FDR-Linking theorem. arXiv:1812.08965, 2018.

Jinjin Tian and Aaditya Ramdas. ADDIS: An adaptive discarding algorithm for online FDR control with conservative nulls. In *Neural Information Processing Systems 32*, 2019.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):822–852, 2022.

Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Ziyu Xu and Aaditya Ramdas. More powerful multiple testing under dependence via randomization. *Bernoulli (forthcoming)*, 2026.

Ziyu Xu, Aldo Solari, Lasse Fischer, Rianne de Heide, Aaditya Ramdas, and Jelle Goeman. Bringing Closure to False Discovery Rate Control: A General Principle for Multiple Testing. arXiv:2509.02517, 2025.

Sonja Zehetmayer, Martin Posch, and Franz Koenig. Online control of the false discovery rate in group-sequential platform trials. *Statistical Methods in Medical Research*, 31(12):2470–2485, 2022.

Yifan Zhang, Zijian Wei, Haojie Ren, and Changliang Zou. e-GAI: e-value-based Generalized α-Investing for Online False Discovery Rate Control. In *International Conference on Machine Learning*, 2026.

Qingyuan Zhao, Dylan S. Small, and Weijie Su. Multiple Testing When Many p-Values are Uniformly Conservative, with Application to Testing Qualitative Interaction in Educational Interventions. *Journal of the American Statistical Association*, 114(527):1291–1304, 2019.

Tijana Zrnic, Daniel Jiang, Aaditya Ramdas, and Michael Jordan. The Power of Batching in Multiple Hypothesis Testing. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Tijana Zrnic, Aaditya Ramdas, and Michael I. Jordan. Asynchronous Online Testing of Multiple Hypotheses. *Journal of Machine Learning Research*, 22(33):1–39, 2021.