We thank all reviewers for their thoughtful reviews. We are grateful that they conclude that "theoretical results of this paper are solid" (R2), the "experiments are conclusive" (R3) and reproducible (R1), and the paper is clear and well written (R1,R2,R3). Below we respond to the reviewer's major concerns and take our responses and the remaining concerns into consideration when revising our paper. Since three reviewers already favor acceptance, we hope that the general skepticism of R4 can be addressed in the discussion phase.

Applicability beyond linear combinations of base test statistics (R1, R3). As we discuss in the conclusion of the submission, specifying the base set of test statistics beforehand is indeed a limitation of our approach. Nevertheless, we have two further comments on this. First, this setting is already flexible enough for many practical applications. One can choose the base test statistics in any creative way as long as it does not depend on the data, e.g., a grid of hyper-parameters, domain-specific kernels, etc. Second, a necessity for our approach is the characterization of the *selection event*. For the considered methods, this is doable because we can solve the optimization problem leading to the optimal test statistic either in closed form or via convex optimization. More general approaches, like continuous optimization of a bandwidth or learning a deep kernel, are not convex problems. To the best of our knowledge, the characterization of the selection event under these scenarios remains an open question.

Motivation and connection to data splitting (R2, R4). It is important to note that we are not suggesting a new test statistic per se. Linear combinations of linear-time MMD estimates have been considered before, e.g., in reference [4] of the submission. For any (normalized) linear projection $\tau_{\boldsymbol{\beta}} = \boldsymbol{\beta}^\top \boldsymbol{\tau} / (\boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta})^{\frac{1}{2}}$ of the base test statistics $\boldsymbol{\tau}$, where $\boldsymbol{\beta}$ is independent of $\boldsymbol{\tau}$, the asymptotic distribution is standard normal (p.3 in the submission). If the independence of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ holds, the asymptotic test power can be derived in closed form for any level α, see Eq.(1) in the submission. To guarantee this independence, previous work used data splitting to estimate the optimal combination $\boldsymbol{\beta}^*$ at the expense of the resulting test power. For us, the motivation is exactly the same, except that we aim to avoid data splitting to achieve higher power (see our experiments). Thus our main contribution is to explicitly correct for the dependence between $\boldsymbol{\beta}^*$ and $\boldsymbol{\tau}$ when the same data is used to estimate both of them. In the revised version, we will make this intuition more understandable, and we will also include a proof sketch of Theorem 1 in the main part.
On the other hand, please note that R1 concludes that the "*proofs look sound*" and that our methods correctly control the Type-I error at the rate α (Fig. 7 in the appendix), which is strong empirical evidence that the "*methodology is sound*," as noted by R3.

Beyond linear-time MMD two-sample tests (R1, R2): Our methods only require asymptotic normality of a set of base test statistics. This is also the case for the B-test (R1) as well as for goodness-of-fit tests (R2) based on the linear-time kernel Stein discrepancy (KSD). However, asymptotic normality under H_0 does not hold for the FSSD given in Jittkritum et al. (2017). Due to the space constraints and to highlight our main contribution, we decided to focus on the simplest case of linear-time MMD estimates and not compare against B-tests or complete U-statistics MMD. We will add a discussion of methods that balance computational efforts with accuracy and the applicability of our methods in the revised version.

Quantification of the advantage over data splitting (R3): The central motivation of our work is indeed that data splitting leads to a decrease in test power. Our main contribution is to show that it is *possible* to learn then conduct the test effectively without resorting to data splitting. In the experiments, we show that our approach outperforms data splitting on all splitting ratios, which was perceived as conclusive by R1,R2,R3. The theoretical characterization of this gain will require partial (if not full) knowledge of *finite-sample* (non-asymptotic) distributions of the statistic under both H_0 and H_1. When the underlying data distribution is unknown (nonparametric setting), these distributions are non-trivial to obtain and remain an open question even in the case of the widely studied T-test statistic. We leave this important direction for future work.

Others: - Computational cost (R4): The computational cost of our approach and data splitting are of the same order. The cost of estimating the test statistic and covariance matrix at once with the entire data set is $\mathcal{O}(n)$, whereas data splitting costs $\mathcal{O}(cn) + \mathcal{O}((1-c)n)$ for $c \in (0, 1)$. Thus at the same computational cost, we achieve higher power as shown in our empirical results.

- "*These [used] kernels are not universal*" (R4): This is a misunderstanding. We actually use various Gaussian kernels for all experiments in Fig. 1 which we report in lines 270-272 of our submission.

- "*How [can one] utilize their own data*" (R1): Thanks for this suggestion. We will update the README to enable an easy integration of new dataset. We are thankful that R1 invested the time and managed to reproduce our results.

- l.140 "The rule for selecting the test statistic from these sets is simply to select the one with the highest value" (R3): Note that we defined $\tau_{\boldsymbol{\beta}}$ including a normalization (thus all $\tau_{\boldsymbol{\beta}}$ have unit variance). Hence, the SNR coincides with the value of the test statistic and determines the power. Thus selecting the one with highest value is sufficient. Also note that the norm constraint on $\boldsymbol{\beta}$ is only to ensure a unique solution, see line 121 in the submission.