We thank the reviewers for their careful reading and constructive comments. Below is a list of the actions we will take to A) clarify certain aspects of the method description (R1/R4), B) present complementary experimental results (R2/R3) and C) update the Python code shared in the supplementary material to solve the packaging issue (R2).

A) Clarifications. They concern statistical assumptions and contributions, as well as take-away from experiments. Improvements to the "flowing logic" (R4) clarification of the "take-away" (R1) in experiments (Sec. 3). We will add an introductory paragraph before Sec. 3.1, presenting briefly (5 lines) the sequence of experiments and their rationale. It amounts to: 1/point spread estimates which are common practice for M/EEG source localization, 2/realistic simulations to show δ-FWER control and benchmark against competing methods, 3/results on real data.

ecd-MTL experimental benefits (R1/R4). ecd-MTL is our privileged solution as it remains competitive in an adversarial simulation (Exp. 1, line 192-197), has best recovery properties in realistic simulations (Exp. 2), is the only method that offers statistical control (Exp. 2), and produces statistics with universal threshold (Exp. 3) contrary to sLORETA.

Statistical guarantees thanks to randomized clustering (R1/R4). Clustering improves conditioning, which allows to verify the Restricted eigenvalue (RE) property (assumptions A1 and A3). RE property is a standard technical assumption (Bickel et al. 2009). The less clusters (i.e. small C), the more A3 is likely to be verified for Prop. 2.2 and 2.3 to hold. Also, with smaller C the sensitivity of ecd-MTL is improved. However, a small C also requires higher spatial tolerance. We then hit a fundamental trade-off for statistical inference between sensitivity and spatial specificity. This is a key contribution of our work. Taking $C = 1000$ seems for the present use case an adequate trade-off to ensure δ-FWER control with reasonable spatial tolerance. Note that the clusters are obtained a priori (cf. Algo. 3). The clusters are not optimized using the target data, but sampled from a set of good candidate parcellations. A key idea of ecdl-MTL is indeed to randomize the definition of the clusters to mitigate the bias due to a specific choice of clusters. Good clusterings are necessary to minimize the compression loss (assumption A2) while having highly variable clustering solutions is necessary to benefit from ensembling. Altogether these considerations motivate the use of the ecd-MTL approach over the d-MTL or the sLORETA methods that do not offer the same statistical guarantees (R1/R4).

Combining heterogeneous sensors (R2/R3/R4). Mixing different types of sensors would violate our modelling assumptions both on temporal correlations and on spatial correlations. A possibility to handle heterogeneous sensors is to follow Massias et al. (2018): Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression, but for the temporal part further developments are required and left for future work.

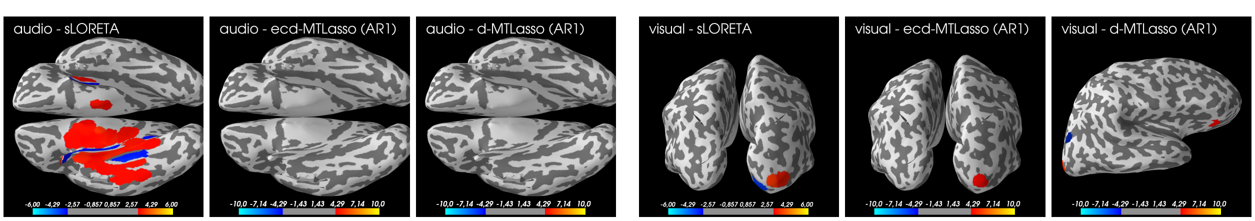
Testing assumptions (R4). Another limitation is the fact that assumptions A1, A2 and A3 are hard to test in practice.

"3.4 Summary, guidelines, limitations" (R1/R4). We will add a sub-section to summarize the aspects discussed above.

Theoretical novelty (R1/R4). Our work is complementary to Mitra et al. (2016): taking the Multi-Task approach allows for a i/ simple statistic test formula, ii/ the integration of auto-correlated noise and iii/ a simplification of mathematical assumptions since they reduce to the RE assumption. Additionally, our present work can be used for solving inverse reconstruction problem of spatially structured data (medical imaging, genomics, geosciences, etc.) (R4).

B) New EEG experiments (R2/R3). Below, we present results on real data keeping only EEG sensors (added to app.). Bilateral auditory activations have historically been hard to infer with EEG sensors: In F1-F3, sLORETA produces only false discoveries (FD) while ecd-MTL and d-MTL make no discoveries. In the visual experiment G1-G3: sLORETA and ecd-MTL produce expected patterns, d-MTL produces expected patterns + one false discovery in the frontal lobe. In the paper, we have emphasized MEG experiments: with more sensors w.r.t. EEG, statistical power is improved (R4).

C) Algorithms and Code. (R2/R4) Our attached code runs fine with sklearn 23.1, but requires running 'celer 0.5dev' (requirements were in the read.md file). Also, we noticed that 'pyvista', which is used for 3D visualization, has evolved and now requires 'pyvistaqt', to use 'mayavi' instead please comment l.90 of the main script. Code will be released and tested on GitHub. Concerning algorithmic clarity (R1/R4), we remind that Alg. 3 in app. synthesizes ecd-MTL. For further clarity about the successive algorithmic extensions, we will add an overview diagram of the methods (R1).



(F1) (F2) (F3) (G1) (G2) (G3)