Many thanks for the helpful comments. We appreciate the praise for the "Extremely good and unique empirical contributions", "Very nice theoretical results" and "novel and sound" in the two high-confidence reviews recommending acceptance, as well as the constructive criticisms, which we will address below.

Markov blanket, Lasso and comparison with other methods: We tried LASSO prior to this work, but the results were not neurophysiologically meaningful. This is understandable in retrospect since causal feature selection via Lasso or Markov Blanket (MB) requires causal sufficiency, let alone curse of dimensionality. Furthermore, with high dimensional data, any algorithm using CI tests has to condition on large variable sets, in which case CI testing is hard and cannot be trusted unless sample sizes are huge. Finally, even if causal sufficiency were to hold, the known MB detection algorithms and Lasso do not *detect* variables but rank them, and gradually evaluate the prediction accuracy by including more variables, according to the ranked order the algorithm returned. This requires a heuristic *hyperparameter* to define what is the right acceptable number of variables to be included in the MB, which both affects the FP and FN and does not provide a straight forward metric (FP, FN) to compare with our method. For completeness, however, below we provide comparison results (to be included in the final version) of our method against available algorithms (average for 10 random graphs): HSIC Lasso (Yamada, 2014), Backwards elimination (BE) with HSIC, and Forward selection (FS) with HSIC for MB discovery (Song, 2007). Lacking space, we selected to examine the most optimistic for the other algorithms case, that of large sample size (800) and two cases of small (20) and large graphs (125 nodes), for sparse (0.2) and dense (0.5, more true causes) edges. We report the % of FP and FN in the number of variables. In sparse large graphs FS gives more FP. Lasso and FS give more FP in small sparse and dense graphs. BE performs worse in small sparse graphs. Overall, our method manages to keep FPs very low ($\sim 2.1\%$) for all dense/sparse, small/large graphs, while other algos' performance varies with the case. Optimal parameters based on the true number of causes was selected for Lasso. BE and FS computations took significantly long. Furthermore, we stress that in these simulations no hidden variables exist, which is an extra advantage for the compared algorithms.

(nodes, sparse)	(20,.2)	(20,.5)	(125,.2)	(125,.5)	(20,.2)	(20,.5)	(125,.2)	(125,.5)	(20,.2)	(20,.5)	(125,.2)	(125,.5)	(20,.2)	(20,.5)	(125,.2)	(125,.5)
	Our method				**Hsic Lasso**				**BE hsic**				**FS hsic**			
FP(%) /FN(%)	**3.5**/31.5	**2**/80	**2.9**/70.3	**0**/80.8	9.5/22.5	5.5/47.5	1.1/77.4	0/84.8	11/23	1.5/79	1.4/77.9	0/97.6	6/25	7.5/26	7.8/47.4	1.1/14.5

Data will be made public /lack of ground truth: The reason why there is no baseline comparison for our EEG results is that this EEG data have not been used for causal inference analysis before, as we ourselves recorded it for this paper. We will indeed make it public upon acceptance, to contribute this fascinating dataset to inspire further causality research in this field that depends on it. Lack of ground truth is a common problem in brain datasets, which is why we compare to conclusions and findings in the literature. Our findings are in accordance with established neuroscientific conclusions about the brain rhythms present during movement in different conditions and we thus believe they are meaningful.

Sufficient but not necessary: The fact that our conditions are sufficient but not necessary may potentially lead to fewer detected causes. Indeed, an empirical example of this was given in section A2, fig. 5 of suppl. submitted alongside the paper, showing a case where both direct causes M_1 and M_2 of target R are rejected. In this example, P_1 and R are not d-separated by M_1 because M_1 is a collider. Moreover, P_2 and R are not d-separated by M_2 due to the path including P_1 and M_1. In this example there are instantaneous effects between the P and the M stages of the causes. This leads to rejecting both causes. This is a counterexample where although the variables are causes of the target, our conditions are not met; thus the ($\leftarrow$) direction of our theorem cannot be proved, thus sufficient but not necessary.

Further explanation on the False Negatives fig. 2: Our method detects direct and indirect causes. If for example $A \to B \to C \to D$ and $E \to D$ in the same graph, our method could identify as causes of D for instance E (direct) and A (indirect). In that case, B and C will be counted as FN, because they were not identified. However, in reality, this is not a problem, because we correctly identified A which is a cause of D (as well as of B and C), and so if we intervene on A we will affect D, which is our ultimate purpose, supposing i.e. A, B, C, E are brain regions and D is the arm speed. Therefore, the number of FN (suppl. fig.9) appears inflated because we consider as causes both the direct and the indirect ones. In case only the direct cause is identified, then its ancestors (indirect causes) will be counted as FN. That is why the number of FN increases with the number of features n and the density of the graph. We stress that the reason why we care more about the FP, is because we address causal problems where a false rejection is less harmful than a false acceptance. If a brain area is falsely identified as a target for stimulation it can be harmful, compared to the harmless case that not all areas are identified. There is no free lunch, and this is a small price to pay to get the linear computation time and the statistical significance of our CI tests with one targeted conditioning variable.

Improvements upon previous methods: 1. To the best of our knowledge, this is the first constraint based algorithm that scales linearly with the number of variables. Previous methods based on CI tests grow exponentially in time with the number of variables, (if sparse data then they grow polynomially), as they require more than one CI test per variable. Therefore, we greatly reduce the computational complexity. 2. Our algorithm builds on tests that condition on only one variable each; previous methods require conditioning on many variables. With this improvement, the statistical strength of our inference is superior compared to cases where there is more than one conditioning variable. Furthermore, due to this improvement, as *reviewer #2* also pointed out, we require a weaker notion of faithfulness. 3. Our method does not assume causal sufficiency - a common assumption which is, however, often violated in real datasets. 4. Finally, although originally for completeness we assume i.i.d. samples, we prove in the suppl. that our method is robust against false positives when the i.i.d. assumption is violated (common violation in real data).