nature portfolio

Peer Review File

Controlling the human connectome with spatially diffuse input signals

Corresponding Author: Professor Richard Betzel

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper extension current optimal control strategies by including distributed stimulation across adjacent network nodes and observes characteristics of targeted node clusters that can more easily move the system to desired brain states.

Overall, this is an important contribution both in studying more realistic neuromodulation scenarios where multiple nodes are affected and in reducing the amount of energy that is needed to attain the desired brain states. There are some points that would need a clarification.

Major points:
- Introduction: it would be good to add focused ultrasound as another Neuromodulation technology. Even though that technology could be very focused (one mm^3) with standard equipment, the focus is more than a cm^3 which could, especially for small subcortical targets, activate multiple network nodes.
- Another point in addition to distance is the myelination of fibres. For example, distant regions that are connected over myelinated fibre tracts might have a faster impact on a region than a nearby region that is linked over unmyelinated fibres. This is something that could be included in future work.
- The authors already mention that gyrification as an important consideration as spatially nearby nodes can be further away considering the distance across the cortical surface. On a positive note, they might highlight that their current approach works straight away for lissencephalic brains (e.g. studies in marmoset monkeys or rodents).

Minor points:
- abstract, first line: remove ‘dro’ at the start
- page 6, right column, ‘the the’ -> ‘the’
- page 17, equation 20: What is the unit for the distance D (mm or cm)?
- Figure S10: grey lines/numbers are hardly visible in a printout.

Reviewer #2

(Remarks to the Author)
general comments
How to control brain dynamics via endogenous state shift or external brain stimulation is a novel and intriguing interdisciplinary topic in the field, which recently has encouraged efforts to combine control theory and brain stimulation studies. The authors raise a new strategy in this manuscript --- instead of precisely controlling the input of each node (local strategy), which is impractical in terms of brain stimulation practice and imprecise for not considering anatomical connections and signal diffusion, they applied an exponential kernel to 'extend' and overlay the impact of each node to their neighbors (spatial strategy). The key finding is that when the optimal control framework is applied to estimate the required energy to reach the target brain state, using the spatial strategy requires less energy than using the local strategy. The authors further analyzed and compared two types of control inputs in this manuscript: an Nx1 whole-brain input, i.e., all brain nodes receive a control signal, and a compressed low-dimensional input, which only targets the cluster centroids to mostly energy-efficiently tweak brain states. The latter approach also aligns nicely with the brain annotations, justified by the biophysical nature of the key brain regions. It may also shed light on future brain stimulation practices. In general, the idea is novel, highly valuable, well-explored and developed in this manuscript, which I believe will bring new insights to the field.

The writing is, however, not so clear in terms of concepts and technical details. I list my comments and questions below, which could be improved for the general readership, including myself.

Abstract and Introduction
1. Many branches/aspects of control theory have been applied to understand features of brain activity and neural plasticity for stability/robustness, energy efficiency, etc. As a major theoretical basis of this study, however, it needs to be clarified why optimal control framework is of great interest here. The authors have dropped many papers using sentences like 'Optimal control has been used to understand the link between brain structure and function...', but it would be more helpful to elaborate on the key findings to lay the ground for this study.
2. Brain state and state transition are the main objects for the optimal control mechanism addressed in this manuscript. Three short paragraphs were used to introduce these concepts, but with the information being scattered and truncated, it is a bit hard for the readers to form a comprehensive picture here. Fig 1 is nicely prepared, so I suggest fine-tuning the introduction section to better center the information around Fig 1.
3. minor: Fig1 caption: is the length of the dash correct here?

results (plus some comments relevant to the M&M section)
1. If I understand it correctly, for detecting brain states, the authors calculated the RMS of each frame's activity for each participant and then selected the frame by the peak RMS of every three neighboring frames for further analysis. So N_peak is the total selected frames across all participants? It was written as 'pattern in total' in the main text before 'pattern' was detected by the further detection algorithm.
2. Then, by calculating the concordance correlation coefficient (CCC) of activities for each pair of frames in the selected N_peak frames, they obtained the N_peak x N_peak matrix filled with CCC values. I suggest adding more explanation when introducing the concordance correlation coefficient in Equation 1. The current version is more like a general description (with two vectors, blah blah) but not tailored for this study.
3. Then, the final pattern was selected by doing a cluster analysis of the matrix of CCC values. Judging by the reference number in the M&M section, the concordance matrix is used in former studies, while the pattern detection methods and Louvain algorithm have not been used a lot before, which I assume is novel in this study. It is also critical for the entire study. I suggest that the authors rewrite this section with supplemental figures to facilitate the understanding of the key methods.
4. one question/concern out of curiosity: The current pattern selection method seems systematically biased in selecting patterns with higher total activity by selecting peak RMS values from all the frames. Is it possible that frames with more silent activity, which can also be called a pattern by definition, are overlooked?
5. A following comment is, what are the input signals like in order to tweak the brain state? Are they all positive values representing brain stimulation that activates the nodes, or a combination of positive and negative values? I didn't see such information in this manuscript.
6. A third question about the input signal is the difference between the whole-brain and low-dimensional inputs. It is not clear to me if the whole-brain input is a Nx1 one-shot input at one discrete time point while the low-dimensional input is a time series.
7. The connectome determines the brain state and input signals. It needs to be better justified why the authors calculated an averaged connectome across all participants and used this one for all the following studies. I can understand if that is for saving computing time, and it seems acceptable that the authors showed that they replicated some findings in another dataset, too. I would find it more convincing if the author can choose one or two individual(s) and examine if the same conclusion holds to achieve the brain state transition based on their own connectome.
8. it is not clear to me why in most of the matrices/heatmaps, the brain state at x-axis is not arranged from 1 to 11.
9. Brain states and canonical resting-state networks are compared in a few figures, but the reasons for doing so and what has been learned from it are not discussed.
10. What is the dimension for Fig4 a,b,c? What are the values on their y-axis? something grouped by the brain region?
11. fig 1 f, the red curve is not defined.
12. The insights for brain stimulation can be extended in the current discussion section, as it is mentioned as one major motivation in the abstract and introduction section.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I have reviewed the changes to the text and the answers to my questions/comments. Apart from the merits I listed in the last round of the reviewing process, the current version is indeed much readable and clear. I enjoyed and learned a lot in the process. Congratulations to the authors for the achievement and I suggest accepting this manuscript.

Dear Editor and Reviewers,

We wish to thank the editor and all reviewers for their thoughtful comments and suggestions. Our enclosed submission is significantly improved as a result. We provide, here, a short summary of those changes:

1. We estimated brain states using an alternative definition. We found that this did not change our conclusions.
2. We estimated control energies for subject-specific connectomes. Again, we found similar effects at the subject level as we did when using a group-representative connectome.
3. We updated large portions of the manuscript to improve clarity.

Throughout this letter we refer to comments made by the reviewers in *italics*. Changes made to the text are shown in *blue italics*. When appropriate, we highlight the section name where changes were implemented. In addition to this response letter, we include two versions of the revised manuscript: a ``marked" version that highlights (again, in blue text) all of the changes made since the original submission and a ``clean" version that includes all changes but without any highlights.

On behalf of the authors,

Rick Betzel

Reviewer #1 (Remarks to the Author):

This paper extension current optimal control strategies by including distributed stimulation across adjacent network nodes and observes characteristics of targeted node clusters that can more easily move the system to desired brain states.

Overall, this is an important contribution both in studying more realistic neuromodulation scenarios where multiple nodes are affected and in reducing the amount of energy that is needed to attain the desired brain states. There are some points that would need a clarification.

We thank the reviewer for their overall positive evaluation of our manuscript and for the opportunity to address their concerns.

Major point #1

Introduction: it would be good to add focused ultrasound as another Neuromodulation technology. Even though that technology could be very focused (one mm^3) with standard equipment, the focus is more than a cm^3 which could, especially for small subcortical targets, activate multiple network nodes.

This is an excellent point and one that we regrettably overlooked. We have updated the manuscript to include focused ultrasound among the additional neuromodulation tools.

The revised text now reads:

> *"This question is presently highly relevant, given the enormous interest in technologies for manipulating ongoing brain activity, including transcranial magnetic stimulation [11], direct current stimulation [12], focused ultrasound [13], as well as chemogenetic applications [14] and optogenetic stimulation [15]."*

Major point #2

Another point in addition to distance is the myelination of fibres. For example, distant regions that are connected over myelinated fibre tracts might have a faster impact on a region than a nearby region that is linked over unmyelinated fibres. This is something that could be included in future work.

This is another important point that was omitted from our discussion. We agree with the reviewer that there is likely regional heterogeneity in terms of spatial effects (myelinated status being one factor that could explain this variability).

We have now updated our "Future Directions" section to discuss this explicitly. It now reads:

> *This study presents a number of opportunities for future studies. One such extension involves the decision to model the spatial dependence of the input signals as a decaying exponential; the framework presented here is generic and amenable to many different spatial relationships. For instance, it would be straightforward to model the effect of space as a Gaussian distribution (centered on a given control site) or a "Mexican hat"' function that depends on spatial proximity. "Additionally, although not explored here, it would be potentially impactful to model heterogeneity across brain regions in terms of their spatial dependence. For instance, this heterogeneity could be used to account for interregional differences in local myelination status, where short-range myelinated fibers have faster conduction velocities compared to their un-myelinated analogues."*

Major point #3

The authors already mention that gyrification as an important consideration as spatially nearby nodes can be further away considering the distance across the cortical surface. On a positive note, they might highlight that their current approach works straight away for lissencephalic brains (e.g. studies in marmoset monkeys or rodents).

Again, the reviewer brings up an excellent point worth noting. We have updated the manuscript to include this point. It now reads:

> *"However, if we imagine that the diffusivity reflects volume conduction, then our distance metric should take into account cortical folding patterns. Accordingly, a more appropriate distance metric would be the geodesic distance between regions along the cortical surface (though we note that this issue is less of a concern in lissencephalic brains--e.g. those of rodents or marmosets). We leave this exploration for a future study."*

Minor points:

abstract, first line: remove 'dro' at the start
page 6, right column, 'the the' -> 'the'
page 17, equation 20: What is the unit for the distance D (mm or cm)?
Figure S10: grey lines/numbers are hardly visible in a printout.

We thank the reviewer for catching the first two typographical errors. They have been updated.

The unit for distance is mm. We have updated that in the main text to clarify this.

We have also updated the grey lines and text in Figures S10. In the updated figure they are now much darker.

Reviewer #2 (Remarks to the Author):

general comments

How to control brain dynamics via endogenous state shift or external brain stimulation is a novel and intriguing interdisciplinary topic in the field, which recently has encouraged efforts to combine control theory and brain stimulation studies. The authors raise a new strategy in this manuscript --- instead of precisely controlling the input of each node (local strategy), which is impractical in terms of brain stimulation practice and imprecise for not considering anatomical connections and signal diffusion, they applied an exponential kernel to 'extend' and overlay the impact of each node to their neighbors (spatial strategy). The key finding is that when the optimal control framework is applied to estimate the required energy to reach the target brain state, using the spatial strategy requires less energy than using the local strategy. The authors further analyzed and compared two types of control inputs in this manuscript: an Nx1 whole-brain input, i.e., all brain nodes receive a control signal, and a compressed low-dimensional input, which only targets the cluster centroids to mostly energy-efficiently tweak brain states. The latter approach also aligns nicely with the brain annotations, justified by the biophysical nature of the key brain regions. It may also shed light on future brain stimulation practices. In general, the idea is novel, highly valuable, well-explored and developed in this manuscript, which I believe will bring new insights to the field. The writing is, however, not so clear in terms of concepts and technical details. I list my comments and questions below, which could be improved for the general readership, including myself.

We thank the reviewer for their overall positive evaluation and for the opportunity to improve the writing and clarity.

Abstract and Introduction

Comment #1.

Many branches/aspects of control theory have been applied to understand features of brain activity and neural plasticity for stability/robustness, energy efficiency, etc. As a major theoretical basis of this study, however, it needs to be clarified why optimal control framework is of great interest here. The authors have dropped many papers using sentences like 'Optimal control has been used to understand the link between brain structure and function...', but it would be more helpful to elaborate on the key findings to lay the ground for this study.

We apologize for the lack of justification for why optimal control, specifically, is relevant. We provide a detailed explanation here and include a paraphrased version of this passage in the main text.

Our brain's activity pattern—its "state"—is constantly changing. It changes when we perform cognitively demanding tasks, when our brain is perturbed/stimulated, and even spontaneously. These changes can be conceptualized as 'transitions' and are presumed to

be effortful (although the amount of effort may vary depending on context—e.g. the state the brain begins in and the state in which it ends).

Optimal control is a well-established engineering framework that models brain state transitions in terms of a networked dynamical system and input/control signals. The system's dynamics explain how brain states change passively over time in the absence of input. The input signals are used to augment that passive change so that the brain ends up in a desired target state. The amplitude and input locations are solved by the optimal control framework once we specify boundary conditions—i.e. the starting and ending states.

The optimal control framework is also well-positioned to investigate the link between structure and function. Brain states are generally thought of (and estimated from) functional data—i.e. activity recordings like fMRI BOLD data. The dynamics that govern both the passive evolution of brain states as well as the diffusion of input signals through the brain depend explicitly upon the connectome—i.e. structural connectivity/white-matter tracts. Accordingly, we can think of optimal control as a framework for studying brain state transitions by exploring the interplay of structure and function.

We now include a clearer motivation and rationale for why we study optimal control and why it is suitable for studying brain-state transitions. The text shown below was added to the introduction:

> *"One promising framework for understanding the effect of these perturbations is network control theory, a branch of engineering concerned with controlling the behavior of networked dynamical systems [16-21]. The optimal control framework, in particular, is especially well-suited for studying effortful brain state transitions. Given an initial and a target state and a model of network dynamics, the optimal control framework asks: where should input signals be delivered and what should those inputs look like so that at time T, the brain is in the target state. Given these constraints, the minimum energy input signals are estimated automatically. Here, energy is measured as the squared input signal, integrated across time and summed across all brain regions. Accordingly, brain state transitions can be ordered based on their energies. Further, because brain states are estimated from functional imaging data and because system dynamics are constrained by the connectome, we can use the optimal control framework to investigate links between structure and function [22-25].*
>
> *Indeed, applications leveraging optimal control are common. It has been used to study brain network development and executive function [26-28],…"*

Comment #2.

Brain state and state transition are the main objects for the optimal control mechanism addressed in this manuscript. Three short paragraphs were used to introduce these concepts, but with the information being scattered and truncated, it is a bit hard for the readers to form a comprehensive picture here. Fig 1 is nicely prepared, so I suggest fine-tuning the introduction section to better center the information around Fig 1.

We agree that these elements – brain states and transitions – could be emphasized more clearly in the introduction. We have updated the first two paragraphs to better explain these terms and how they relate to optimal control. The revised text reads:

> *"The human connectome is a network map of the brain's physical wiring (Sporns et al 2005; Sporns 2011). Nodes correspond to neural elements -- brain regions at the macroscale -- where computations are carried out locally. The outcomes of those computations are relayed from one node to another along network edges -- fasciculated white-matter (Wandell et al 2016, Jbabdi et al 2015). Accordingly, the structure of the connectome plays an important role in shaping interregional communication (Avena Koenigsberger et al 2018), patterns of brain activity, and its correlation structure--i.e. functional connectivity (FC) (Suarez et al 2020; Fotiadis et al 2024).*
>
> *Brain states are central concepts in neuroscience (Hancock et al 2024, Greene et al 2023). Although there exists a plurality of definitions, they are commonly derived from instantaneous estimates of connectivity or, more frequently, activity (Fig 1a). For instance, at a given time, t, one could assemble the fMRI BOLD magnitude from all N regions into an $N \times 1$ vector. We refer to this multivariate pattern as a brain state, as it specifies the coordinates of the brain in an N-dimensional state space at t (Greene et al 2023). As the brain's multivariate activity pattern fluctuates over time, the brain's location in state space changes, tracing out a trajectory. We refer to this process -- moving from one brain state to another -- as a brain state transition."*

Comment #3.
Fig1 caption: is the length of the dash correct here?

There are several dashes in this caption. For the sake of clarity, we have replaced them with alternative punctuation. Specifically, we replaced "—time-varying inputs—" and "—e.g. a region's center of mass" with ", i.e. time-varying inputs," and ", e.g. a region's center of mass".

Results (plus some comments relevant to the M&M section)
Comment #1.

If I understand it correctly, for detecting brain states, the authors calculated the RMS of each frame's activity for each participant and then selected the frame by the peak RMS of every three neighboring frames for further analysis. So N_peak is the total selected frames across all participants? It was written as 'pattern in total' in the main text before 'pattern' was detected by the further detection algorithm.

We apologize for the confusion and are grateful for the opportunity to clarify.

Here, we calculated RMS for all frames and for every participant. For a given participant's scan, we identified all local maxima as any time point, t, such $RMS(t) > RMS(t-1)$ and $RMS(t) > RMS(t+1)$. For each such local maximum, we retained the whole-brain activity pattern at time t as an $N \times 1$ vector, where N is the number of nodes.

The variable N_{peaks} corresponds to the total number of local maxima identified across all participants.

We have added a clarifying remark in the "Results" section, but also refer the reviewer to the "Materials & Methods" where we describe this procedure in more detail.

The clarifying remark reads:

> *"We define peaks as frames whose RMS was greater than that of both the preceding and following frames. That is, peaks are frames where $RMS(t-1) < RMS(t) > RMS(t+1)$. We then aggregate the corresponding patterns ($N_{peak} = 3443$ patterns in total) across participants… "*

Comment #2.

Then, by calculating the concordance correlation coefficient (CCC) of activities for each pair of frames in the selected N_peak frames, they obtained the N_peak x N_peak matrix filled with CCC values. I suggest adding more explanation when introducing the concordance correlation coefficient in Equation 1. The current version is more like a general description (with two vectors, blah blah) but not tailored for this study.

We apologize for lack of details. Given that the reviewer is referring to "two vectors," a phrase only used in "Materials & Methods", we presume that this comment refers to that section. To better clarify what the concordance coefficient measures and how it relates to the more familiar correlation coefficient, we have added the following statement at the end of the sub-section "Brain State Detection":

> *"The rows and columns of this matrix corresponded to peak activation patterns. The entries in the matrix are concordance coefficients, and can be interpreted as a*

measure of how similar two activity patterns are to one another. Unlike the product-moment correlation coefficient, which performs a standardization procedure (z-scoring to remove differences in mean and variance) prior to estimating the correlation, concordance explicitly takes those metrics into account when estimating similarity. This is important for state estimation; z-scoring activation patterns equates low- and high-amplitude activity (as well as low- and high-variability) by falsely putting them on a comparable scale. Thus concordance allows us to defines states that not only have similar patterns, but also amplitude and variability."

Comment #3.

Then, the final pattern was selected by doing a cluster analysis of the matrix of CCC values. Judging by the reference number in the M&M section, the concordance matrix is used in former studies, while the pattern detection methods and Louvain algorithm have not been used a lot before, which I assume is novel in this study. It is also critical for the entire study. I suggest that the authors rewrite this section with supplemental figures to facilitate the understanding of the key methods.

Again, we apologize for the lack of clarity. Both the concordance *and* the Louvain algorithm for clustering are both used widely. If anything, the Louvain algorithm is the more widely used (the paper introducing the algorithm in 2008 has now been cited almost 24,000 times at the time of writing). We have updated the text in this section, making it clear exactly what the Louvain algorithm is doing along with some additional context.

The revised text reads:

"We use the Louvain algorithm to optimize Q [105]. Briefly, the Louvain algorithm is a greedy, multi-stage algorithm. The first stage consists of a loop through all nodes in the network in random order. At each step in the loop, the current node is merged with all other possible communities and the resulting ΔQ calculated. The merger corresponding to the greatest increase is Q is then performed (this is the component of the algorithm that is considered "greedy"). The loop is repeated until no single-node mergers yield a $\Delta Q > 0$. At this point, the first stage ends and the second stage begins. The second stage is an aggregation stage, in which a new "metanetwork" is constructed based on the community assignments. That is, all nodes assigned a given community label are combined into a single metanode, and the connections between pairs of metanodes are calculated as the total weight of connections between their constituent nodes. Note that the metanetwork will generally exhibit self-loops. Then the algorithm returns to the first stage, repeating this entire procedure until a complete pass--stages one and two--yield no changes in Q.

The Louvain algorithm is stochastic, meaning that repeated runs of the algorithm will typically yield dissimilar estimates of community structure. Accordingly, the Louvain algorithm must be run multiple times and the outputs aggregated into consensus communities [46]. Here, we repeat the Louvain algorithm 1000 times, which yields 1000 estimates of communities. Briefly, our consensus community algorithm works by calculating for every pair of peak activation maps the fraction of the 1000 Louvain runs in which they were assigned to the same community. It also calculates the expected co-assignment probability (by randomly permuting the detected community labels). This allows us to compare the observed and expected coassignment probabilities, setting up a consensus modularity function, which we optimize using the same Louvain algorithm. Clustering the consensus matrix reduces variability across the detected communities, oftentimes yielding a single solution across the 1000 runs. If that is the case, then the algorithm has discovered the "consensus clusters." If variability remains, the consensus algorithm is applied again to the new estimates of communities. These steps -- generating estimates of communities and calculating the coassignment matrix -- are repeated until the estimated clusters are all identical. Note that this specific consensus clustering algorithm has been applied previously [84,102-104]."

Comment #4.

one question/concern out of curiosity: The current pattern selection method seems systematically biased in selecting patterns with higher total activity by selecting peak RMS values from all the frames. Is it possible that frames with more silent activity, which can also be called a pattern by definition, are overlooked?

This is an important observation. To define brain states in the manuscript, we excluded low-amplitude frames, focusing instead on peaks. The rationale for doing this is multifold. First, the BOLD-fMRI signal is strongly autocorrelated. Therefore, sampling every frame is likely "overkill" – temporally proximal frames correspond to, largely, the same activation patterns and are unexpected to wildly change the detected brain states. Second, in previous work (Betzel et al 2022, *Neuroimage*, https://doi.org/10.1016/j.neuroimage.2022.118993) we showed that low-amplitude frames, specifically, are more likely to be contaminated by motion artifacts and contribute little to the specificity of functional connectivity.

Nonetheless, we agree that it would be a useful exercise to confirm that this analysis decision does not obviously impact our results. To this end, we repeated our state detection analysis using all frames (not just peaks) as input to the clustering algorithm. Following post-processing—i.e. dropping centroids that appear infrequently (in fewer than 50% of participants)—we obtained 21 brain states. Although some of these states were novel, likely reflecting low-amplitude frames, we found a close correspondence between those reported in main text (estimated from peaks) and those detected here (estimated using all frames). As a demonstration, we used the "Hungarian algorithm" (Kuhn 1955,

https://doi.org/10.1002/nav.3800020109) to uniquely match states. We found a strong correspondence – the mean similarity, measured here as 'concordance', between matched states was 0.79 ± 0.11 (see Fig. S6a).

Next, we calculated the control energy to transition to and from each of the 21 brain states (a total of 441 transitions). We repeated this analysis using the traditional input strategy, wherein input signals are delivered locally to individual nodes as well as the spatial input strategy. As in the main text, we found that on a per-transition basis, the spatial model required significantly less energy than the local model.

These observations suggest that our results are likely not dependent on our decision to define brain states based on high-amplitude frames and generalize to at least one other reasonable definition of brain state. We include the figure shown below in the supplementary material.

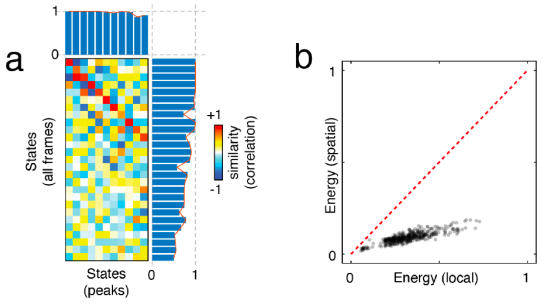


Figure S6. **Replicating main findings using alternative definition of brain states.** In the main text, we identified brain states by clustering activity patterns from peaks in the global activity amplitude (root mean square). Here, we repeated the primary analysis using an alternative definition of brain state. Namely, we clustered all frames, spanning all amplitudes of activity. Using this approach, we found 21 states instead of 11. However, among those 21 states were 11 that closely matched those reported in the main text. (*a*) Correlation (similarity) of states estimated from peak frames with those estimated using all frames. States were matched using the Hungarian algorithm. In the margins, we show on a scale of 0-1, the fraction of participants in which a given state appeared. (*b*) We also calculated the control energy needed to transition between all pairs of states. In line with results reported in the main text, we found that, on a per-transition basis, the energy required using the spatial model was less than what was required using the local model.

Comment #5.

A following comment is, what are the input signals like in order to tweak the brain state? Are they all positive values representing brain stimulation that activates the nodes, or a combination of positive and negative values? I didn't see such information in this manuscript.

We apologize for not highlighting these details. In general, the input signals are signed and continuously weighted (i.e. not discrete values). We now note this explicitly in the "Materials and Methods" section and "Optimal control" sub-section of the manuscript, stating:

"Note also, that in general, the input signals, $\boldsymbol{u}$, are signed and weighted. That is, they are not discrete "on" or "off" signals. See Parkes et al (2023) for more details."

Comment #6.

A third question about the input signal is the difference between the whole-brain and low-dimensional inputs. It is not clear to me if the whole-brain input is a Nx1 one-shot input at one discrete time point while the low-dimensional input is a time series.

We apologize for the confusion. The whole-brain input is not a $N \times 1$ vector. Rather, each of the N nodes receives its own input signal. That is, without the dimension-reduction step, optimal control constructs N totally separate (but possibly correlated) signals. The dimension-reduction step identifies those correlations and if they are sufficiently strong, our algorithm groups them into a single cluster. From that cluster, we estimate a single representative time series (this is the dimension reduction step).

We have updated the manuscript to clarify the dimensionality of these constructs. The revised text reads:

> In the main text we described an algorithm for approximating optimal control using a much smaller set of input signals. The algorithm for doing so proceeds as follows. First, using every node as an input site, we obtain the optimal input signals. We denote the full set of input signals as $\boldsymbol{u}$, which has dimensions $[T \times N]$, where T is the number of time points and N is the number of nodes. We denote the input to node i as $\boldsymbol{u}_i$ (or equivalently, the ith column of $\boldsymbol{u}$). The dimensions of $\boldsymbol{u}_i$ is $[T \times 1]$. We then use the k-means clustering algorithm (varying the number of clusters from $k = 2$ to $k = 40$; Euclidean distance) to partition input signals based on their temporal similarity. We then estimate the mean input signal for each cluster, c:
>
> $$\boldsymbol{u}_c = \frac{1}{N_c} \sum_{i \in c} \boldsymbol{u}_i$$
>
> Note that here, the dimensions of $\boldsymbol{u}_c$ is $[T \times 1]$, even though the size of cluster c, denoted by N_c, is generally much greater.

Comment #7.

The connectome determines the brain state and input signals. It needs to be better justified why the authors calculated an averaged connectome across all participants and used this one for all the following studies. I can understand if that is for saving computing time, and it seems acceptable that the authors showed that they replicated some findings in another dataset, too. I would find it more convincing if the author can choose one or two individual(s) and examine if the same conclusion holds to achieve the brain state transition

based on their own connectome.

This is a fair point. We focused on the group connectome because structural connectivity reconstructed from diffusion MRI and tractography can be noisy and error-prone. Further, the primary contribution of this manuscript is the comparison of two methods for delivering the input signals to the brain and not on individual differences.

Nonetheless, we agree that it would be useful to show that similar effects hold with subject-level connectomes. To this end, we repeated the main analysis using connectome data from three participants chosen randomly. Specifically, we fixed the spatial parameter to $\beta = 0.15$ (the value we focus on in the main text using the group connectome) and repeated all 121 transitions (11 states to 11 states). We show the results below. In general, we find that on a per-transition basis, the spatial model always required less control energy. These observations are directly in line with those reported in the main text. We now include this figure in our supplement.

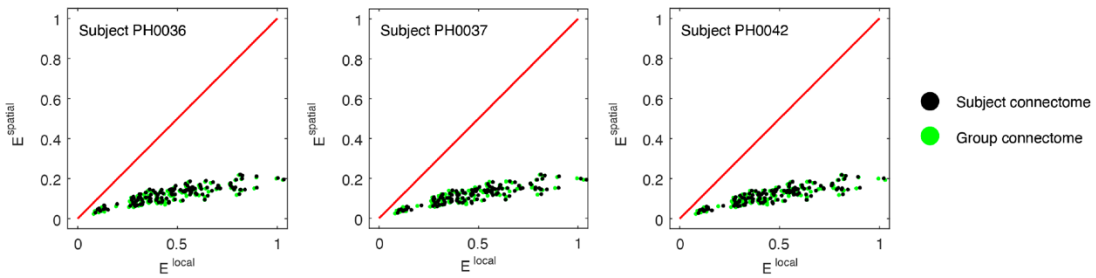


Figure S5. **Replicating main findings using single-subject connectomes.** In the main text, we used a group-representative connectome to estimate control energies. Here, we show that we obtain similar results using connectomes estimated at the subject level. Specifically, we calculate the global control energy for all 121 transitions (11 states to 11 states). In every instance, the spatially informed input strategy resulted in reduced control energy compared to the standard input model.

Comment #8.

it is not clear to me why in most of the matrices/heatmaps, the brain state at x-axis is not arranged from 1 to 11.

We appreciate the reviewer for pointing this out. The numbering of brain states was based on frequency; the state that appeared most often was assigned the label "1." The state that appeared second most was called "2", the third most called "3", and so on.

The ordering in which brain states appear in some of the figure panels was motivated by aesthetics. In Figure 2b, we calculated system-level activations for each brain state. We then ordered the states so that states with similar activation patterns were proximal to one another (using the "brain connectivity toolbox" function "reorder_matrix.m". We felt that this ordering was more aesthetically pleasing and used it throughout the manuscript.

We note, however, that the labelling of states is entirely arbitrary and in does not impact the results. We have updated the text to clarify that the maps are not presented ordinally. The revised figure caption reads:

> *"The color points corresponds to their pairwise similarity from panel e. Note that in panels b-e, brain states are not ordered numerically. Rather, they are ordered so that states with similar activity patterns appear sequentially. This ordering was obtained using the function reorder_matrix.m in the Brain Connectivity Toolbox with the similarity matrix (panel e) as input [Rubinov & Sporns 2010]."*

Comment #9.

Brain states and canonical resting-state networks are compared in a few figures, but the reasons for doing so and what has been learned from it are not discussed.

We apologize for the lack of clarification. Resting-state networks, as the reviewer alludes, are canonical and recognizable to most neuroimagers, cognitive neuroscientists, and network neuroscientists. They serve as familiar reference when contextualizing a brain map—e.g. activation patterns, brain states, and regional control energies.

In other words, the resting-state networks allow us to make contextualize findings and make statements like "Brain state 1 corresponds to increased and decreased activity in the Salience/Ventral attention network and the default mode network, respectively."

Comment #10.

What is the dimension for Fig4 a,b,c? What are the values on their y-axis? something grouped by the brain region?

We apologize for this omission. The dimension of the arrays shown in Fig 4a,b,c is [Node x Target State]. In this specific case, the number of nodes I 1000 and the number of target states is 121 (11 states, so 11^2 possible state transitions). Importantly, the nodes have been ordered by resting-state network (taken from Schaefer et al 2018). The boundaries between networks are shown with dashed lines (although they are admittedly difficult to see). We have made those lines thicker and include a description of the array dimensions and ordering in the figure caption.

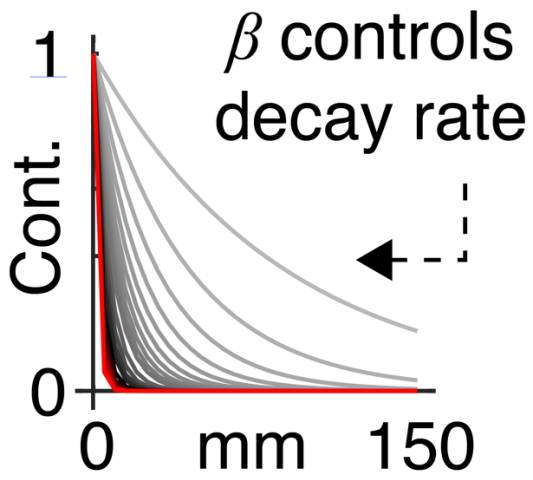
The updated caption reads:

> *"Panels a, b, and c show the "local", "spatial", and "effective" input energies at each brain region and for each of the 121 control tasks, thus each array has dimensions* $[1000 \times 121]$. *Note that the columns in each matrix are ordered based on target*

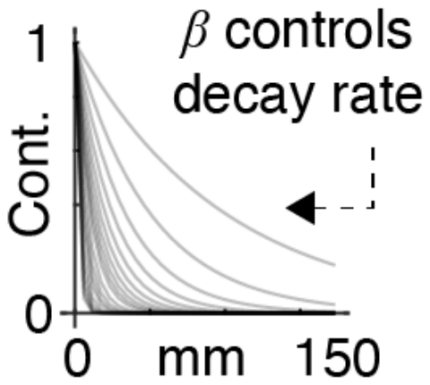
state. The rows have been ordered based on network assignments from [Schaefer et al 2018]. The gray text above the heatmaps indicates …"

Comment #11.
fig 1 f, the red curve is not defined.

We apologize for the omission. Each of curves shown in that plot corresponds to an exponential with a different beta parameter. Though not evident in the image, the curves are mapped to colors—black corresponding to smaller betas and reds corresponding to larger betas. See enlarged figure below.



The red curve is simply the most extreme and, because it is in the foreground, it occludes other exponentials with strong spatial penalties (large beta values). To avoid any confusion, we have simply removed the red-to-black gradient. In the updated figure, shown below, all curves are grayscale with precisely the same intensity.



Comment #12.

The insights for brain stimulation can be extended in the current discussion section, as it is mentioned as one major motivation in the abstract and introduction section.

We appreciate the opportunity expand this section. We have updated the "Discussion" accordingly. The revised text reads:

> *"Selectively modulating brain activity is a grand ambition in neuroscience. Achieving this goal requires not only new stimulation techniques [Krauss et al 2021; McCormick et al 2020], but also developing new models that accurately predict the outcome of stimulation--e.g. given a particular set of stimulation parameters and locations, how is activity throughout the brain disrupted [Horn et al 2020]? Network control theory (and optimal control, specifically), relies on a specific model of whole-brain activity. Although this model is relatively abstract, it nonetheless takes into account the connectome and can readily incorporate inputs, making it potentially well-suited for predicting the effects of stimulation. However, in its standard form, optimal control still relies on unrealistic assumptions. One such assumption is that stimulus energy can be delivered to a single brain region, without diffusing to its spatially adjacent neighbors. The model we study here allows users to flexibly incorporate spatial priors into the optimal control framework. These priors dictate the effective spatial resolution of the neurostimulation apparatus and can, in principle, be tailored to different techniques--e.g. different settings for TMS versus focused ultrasound. In bridging this modelling gap, we enhance the biophysical realism of optimal control models. Important follow-up steps include using the spatially-informed model to make predictions and to test those predictions experimentally. Are there particular sets of stimulation parameters for which the optimal control framework and underlying dynamical model accurately predict their effect? Can the gap between prediction and observation be narrowed by changing the spatial parameters of the optimal control model? These and related questions should be the focus of future work seeking to introduce optimal control in the neurostimulation/modulation domain."*