Coprocessor Actor Critic: A Model-Based Reinforcement Learning Approach For Adaptive Brain Stimulation

Michelle Pan [*1] **Mariah Schrum** [*1] **Vivek Myers** [1] **Erdem Bıyık** [2] **Anca Dragan** [1]

Abstract

Adaptive brain stimulation can treat neurological conditions such as Parkinson's disease and post-stroke motor deficits by influencing abnormal neural activity. Because of patient heterogeneity, each patient requires a unique stimulation policy to achieve optimal neural responses. Model-free reinforcement learning (MFRL) holds promise in learning effective policies for a variety of similar control tasks, but is limited in domains like brain stimulation by a need for numerous costly environment interactions. In this work we introduce Coprocessor Actor Critic, a novel, model-based reinforcement learning (MBRL) approach for learning neural coprocessor policies for brain stimulation. Our key insight is that coprocessor policy learning is a combination of learning how to act optimally in the world and learning how to induce optimal actions in the world through stimulation of an injured brain. We show that our approach overcomes the limitations of traditional MFRL methods in terms of sample efficiency and task success and outperforms baseline MBRL approaches in a neurologically realistic model of an injured brain.

1. Introduction

A neural coprocessor is a form of brain-computer interface (BCI) that can transmit signals to and from the brain (Rao, 2019; Oehrn et al., 2023). These interfaces can be used to treat a variety of neurological conditions by influencing abnormal neural activity (Lozano et al., 2019; Little et al., 2013b). In patients who suffer from conditions such as Parkinson's disease and dystonia, brain stimulation has the ability to steer neural activity towards activity regions which manifest in reduced disease symptoms (Hu & Stead, 2014; Groiss et al., 2009). Adaptive brain stimulation can be employed not just to guide neural activity towards specific activity patterns but can also aid impaired patients in accomplishing external task objectives (Bryan et al., 2022; Cunha et al., 2015). Stroke patients are one patient population that can benefit from this aspect of brain stimulation. Stroke patients suffer injury to the brain that often results in loss of motor control and an inability to complete basic tasks, such as reaching for and grasping an object (Hatem et al., 2016). Stroke patients often struggle with these seemingly simple motor tasks due to stroke-induced lesions in the brain that can interrupt the propagation of neurological signals within and between cortical modules (Ingwersen et al., 2021; Choi et al., 2023). Due to the resultant motor deficits, a patient may recognize the target location and intend to move their arm to the perceived position, but struggle to do so (Choi et al., 2023). Adaptive brain stimulation exhibits potential for reducing motor impairment and restoring lost function via adaptive stimulation, enabling these patients to operate more effectively in the world (Elias et al., 2018; Ganguly et al., 2022).

In this work, we investigate the development of neural coprocessors to deliver adaptive brain stimulation for rehabilitation using an *in silico* model of brain injury. Neural coprocessors rely on artificial intelligence techniques to learn a brain stimulation policy that appropriately shapes neural activity based upon the current state of the patient (Rao, 2019). An effective coprocessor policy can compensate for lost mobility and paretic motor deficits post-stroke (Bryan et al., 2022). However, there are several challenges in developing an effective coprocessor policy. Because of the heterogeneity of patients' brains, their disease manifestation, and the location of the stimuli, the optimal coprocessor policy is unique for each patient (Visanji et al., 2022). Furthermore, due to the complexity of the brain, closed-loop coprocessor policies are difficult to hand-engineer (Oehrn et al., 2023). Instead, individualized coprocessor policies may be learned through interaction with the patient to ensure that the coprocessor aligns seamlessly with the unique

[*]Equal contribution [1]Department of Electrical Engineering and Computer Sciences, UC Berkeley [2]Department of Computer Science, University of Southern California. Correspondence to: Michelle Pan <michellepan@berkeley.edu>.

We make our code public at https://github.com/michelllepan/neural-coprocessors.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

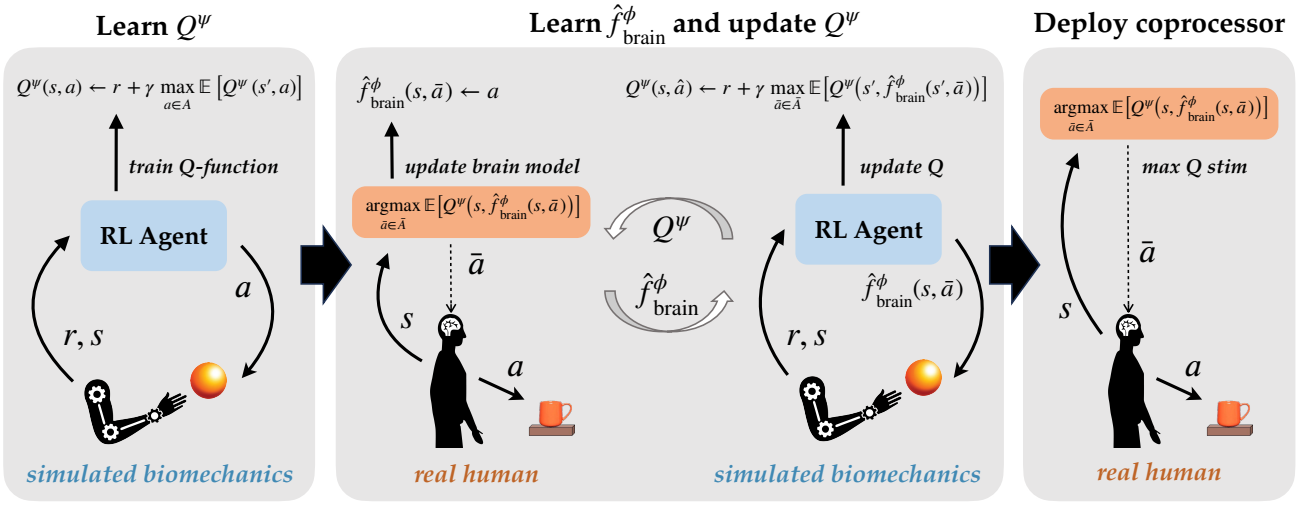

Figure 1. Overview of our framework. We first learn the Q-function, Q^{ψ}, for world actions, a, via a biomechanically realistic simulator. We then learn the mapping, $\widehat{F}^{\phi}_{\text{brain}}$, from coprocessor actions, $\bar{a}$, to world actions. Simultaneously, we update Q^{ψ} to account for the altered MDP.

characteristics of the patient's brain and condition. Due to the necessity for online learning, adaptive brain stimulation poses a compelling paradigm which can benefit from the application of reinforcement learning (RL) techniques.

Model-free reinforcement learning (MFRL) has shown promise for learning high-quality policies in many different control tasks that resemble brain stimulation (Huang, 2020). However, MFRL often requires numerous environment interactions to learn a sufficiently good policy. When stimulating the brain, interactions with patients are costly, and thus, this domain requires an approach which can quickly learn a policy with few patient interactions. Moreover, inappropriate stimulation can produce negative side-effects such as cognitive disturbances, dyskinesia, and mood changes, further necessitating efficient learning algorithms (Ashmaig et al., 2018; Buhmann et al., 2017). To this end, we introduce a novel model-based reinforcement learning (MBRL) approach for coprocessor stimulation that outperforms state-of-the-art MFRL and MBRL approaches in terms of both sample efficiency and task success.

Our key insight is that we can minimize online patient interaction by breaking coprocessor policy learning into two phases: 1) learning how to act optimally in the world, and 2) learning how to achieve optimal world actions via brain stimulation. With access to a biomechanically realistic simulator, we can learn the former without any patient interaction, enabling us to focus online interactions on learning the mapping from coprocessor stimulations to world actions. By separating these two components, we are able to improve the sample efficiency and performance of the coprocessor. An overview of our approach is presented in Figure 1. We define world actions as the tangible movements performed by the patient in their environment, distinguishing them from actions (i.e., stimulations) initiated by the coprocessor.

To learn how to act optimally in the world, we rely on a biomechanical simulator. Various physics based simulators (e.g., Caggiano et al. (2022) and Delp et al. (2007)) have been developed that enable the physiologically realistic simulation of human biomechanics. These simulations are capable of modeling both complex human physiology and dynamic environment interactions, such as the complexities encountered in our example reaching task. We leverage a biomechanical simulator to learn the value of a human executing a world action (e.g., moving their arm) in a given state (e.g., arm joint positions), for a given task objective (e.g., reaching a target location). That is, we employ a simulator to learn the Q-function for world actions.

While the mechanisms involved in executing a world action will be similar for every individual and can therefore be reliably modeled in simulation, because of the complexity and heterogeneity of the human brain, the effect of coprocessor stimulation cannot be as readily predicted. Instead, we must learn how to achieve optimal world actions via stimulation in an online fashion. To efficiently learn how coprocessor stimulations map onto world actions, we leverage the simulator-derived Q-function to guide stimulation sampling during online learning, thus focusing model learning on high-value regions of the world-action space. Because not all world actions will necessarily be realizable by the injured brain, the learned Q-function may be over-optimistic. To solve this problem, we iteratively update the Q-function as we learn the brain model. We demonstrate that this method is orders of magnitude more efficient than learning a stimulation policy from scratch via standard MFRL. Once this

mapping is learned, our coprocessor policy selects stimulations that produce world actions of maximum value, as defined by the simulator-derived Q-function.

We evaluate our approach in standard continuous control tasks as well as a novel, physiologically and neurologically realistic stroke domain. We demonstrate that our approach learns a better policy in fewer interactions compared to baseline MFRL and MBRL approaches. Additionally, we show that our approach is able to better aid the patient in accomplishing a given task during the online learning phase compared to baselines. With this work, we hope to set the groundwork for implementing RL solutions for adaptive brain stimulation and to pave the way for RL researchers to further study this problem. We contribute the following:

1. A novel, model-based RL approach for learning a neural coprocessor policy for closed-loop adaptive brain stimulation.
2. A physiologically and neurologically realistic RL benchmark environment for adaptive brain stimulation for stroke-relevant tasks.
3. Results showing improved sample efficiency and higher training and evaluation reward relative to MFRL and MBRL baselines.

2. Related Work

Current clinical applications of brain stimulation are open loop, i.e., non-adaptive (Khanna et al., 2021; Ganguly et al., 2022; Lu et al., 2020b). For instance, when selecting deep brain stimulation (DBS) parameters for treating Parkinson's in clinical settings, the surgeon performs a trial and error search to determine the set of stimulation parameters that best treat patient symptoms (Ghasemi et al., 2018). However, propelled by advances in medical technology and a better understanding of neurophysiological signals, closed-loop coprocessors are emerging as a new and promising paradigm (Frey et al., 2022). Closed-loop DBS relies on feedback signals either in the form of decoded brain readings or other external state information to dynamically adjust stimulation and deliver more precise interventions for patients. One challenge in closed-loop brain stimulation is the formulation of an effective control policy. Various methods have been proposed to develop adaptive DBS policies based on patient state information (Little et al., 2013a; Bronte-Stewart et al., 2020; Oehrn et al., 2023). For instance, Little et al. (2013a) proposed an approach that modulates stimulation parameters based upon a user-defined threshold of local evoked potentials. However, these manually crafted strategies often fall short of capturing the intricate interplay between DBS and the brain, and are difficult to personalize for individual patients.

Reinforcement Learning for Brain Stimulation: An alternative to hand-engineering policies is to leverage reinforcement learning techniques and learn a stimulation policy via data collected through patient interactions (Schrum et al., 2022; Gao et al., 2023; Coventry & Bartlett, 2023). Several prior works have explored RL methods for closed-loop brain stimulation (Lu et al., 2020a; Gao et al., 2023). Gao et al. (2023) investigated employing a MFRL approach with offline warm-starting to learn an effective stimulation policy for Parkinson's patients. Similarly, Lu et al. (2020a) proposed an actor-critic method and demonstrated its performance in simulation. However, such MFRL methods require a large number of patient interactions and thus, while effective in simulation, MFRL is difficult to deploy in the real world (Dulac-Arnold et al., 2021).

Model Based Reinforcement Learning: An alternative to MFRL is Model-Based Reinforcement Learning (MBRL). MBRL reduces training time and improves learning efficiency by utilizing a predictive dynamics model to learn an effective policy (Valencia et al., 2023; Janner et al., 2021). MBRL can typically be broken down into two steps: 1) dynamic model learning followed by 2) integration and planning (Moerland et al., 2022). The downside of MBRL is that the model used for planning may be inaccurate, thus producing suboptimal plans (Abbeel et al., 2006). Our approach minimizes this risk by leveraging a simulation-derived Q-function to guide model learning and sample stimulations that produce high-value environment actions, enabling us to robustly and efficiently learn the brain dynamics model.

3. Problem Setup

We consider a world MDP with continuous state and action spaces defined by the tuple ($\mathcal{S}$, $\mathcal{A}$, P, R). $\mathcal{S}$ defines the state space and $\mathcal{A}$ the world action space. P: $\mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$ denotes the probability distribution of the next state when action a is taken at state s. $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, defines the reward function.

Using this MDP to describe the world, we construct a second MDP from the perspective of the coprocessor stimulations. We assume there is some (potentially nondeterministic) mapping $F_{\text{brain}} : S \times \bar{\mathcal{A}} \to \mathcal{A}$ that converts stimulations to the resulting world actions in a state-dependent fashion. We then define the augmented coprocessor MDP, $\overline{M} = (\mathcal{S}, \bar{\mathcal{A}}, \overline{P}, R)$ where $\bar{\mathcal{A}}$ is the space of possible coprocessor stimulation actions and the probability distribution over the next state is defined in Equation (1).

$$\overline{P}(s' \mid s, \bar{a}) = \mathbb{E}\Big[P\big(s' \mid s, F_{\text{brain}}(s, \bar{a})\big)\Big]. \tag{1}$$

In Equation (1) the expectation is over the stochasticity of F_{brain}.

The objective is to learn a policy $\pi(\bar{a}_t \mid s_t)$ that maximizes the task reward in the coprocessor MDP, $\overline{M}$. In simulation we do not have access to the true brain model F_{brain}, so we use a learned model $\widehat{F}^{\phi}_{\text{brain}}$, yielding a simulated version of the coprocessor MDP, $\widehat{M} = (\mathcal{S}, \bar{\mathcal{A}}, \widehat{P}, R)$:

$$\widehat{P}(s' \mid s, \bar{a}) = \mathbb{E}\Big[P\big(s' \mid s, \widehat{F}^{\phi}_{\text{brain}}(s, \bar{a})\big)\Big]. \tag{2}$$

4. Methodology

Algorithm 1 Coprocessor Actor Critic (CopAC)

Require: world MDP $M = (\mathcal{S}, \mathcal{A}, P, R)$, $s_0 \sim p_0$
Require: stimulation space $\bar{\mathcal{A}}$, injured brain F_{brain}
1: initialize $\widehat{F}^{\phi}_{\text{brain}}$
2: **while** access to simulator **do**
3: rollout π in M with experiences (s, a, r, s')
4: fit $Q^{\psi}(s, a)$ with Equation (4)
5: **end while**
6: **while** access to injured brain **do**
7: rollout $\bar{\pi}$ in $\overline{M}$ with experiences $(s, \bar{a}, r, s')$
8: $a \leftarrow F_{\text{brain}}(s, \bar{a})$
9: fit $\widehat{F}^{\phi}_{\text{brain}}(s, \bar{a})$ to a
10: **while** not converged **do**
11: rollout $\bar{\pi}$ in $\widehat{M}$ with experiences $(s, \bar{a}, r, s')$
12: $\hat{a} \leftarrow \widehat{F}^{\phi}_{\text{brain}}(s, \bar{a})$
13: fit $Q^{\psi}(s, \hat{a})$ using Equation (6)
14: **end while**
15: **end while**
16: **return** $\bar{\pi}$

We now present our approach for efficiently learning a patient-specific coprocessor policy. Our key insight is that we can separate the policy learning into learning the value of world actions followed by learning to produce high-value actions through stimulation.

Our approach is detailed in Algorithm 1 and consists of three steps: 1) training a world-action value model, Q^{ψ}, 2) training the brain model, $\widehat{F}^{\phi}_{\text{brain}}$, and 3) updating the world-action value model, Q^{ψ}. We alternate between steps 2 and 3 during online patient interaction.

4.1. Training world-action value model

Our goal is to leverage a biomechanical simulator to simulate the effect of a world action a on a world state s, given the world MDP, M. Via this simulation, we can learn a world policy π for how to act optimally in the world without having to directly interact with the patient. We assume that the optimal world policy is consistent across patients (e.g., though their neural activity may differ, patients will reach the same target object by taking the same world actions) and can be readily simulated via the biomechanical simulator. We leverage this simulator to derive the Q-function, Q^{ψ} for the optimal world policy, π (Algorithm 1 lines 2-5). The world policy and Q-update are defined in Equation (3) and Equation (4) respectively.

$$\pi(s) \triangleq \arg\max_{a \in \mathcal{A}} \mathbb{E}\big[Q^{\psi}(s, a)\big] \tag{3}$$

$$Q^{\psi}(s, a) \leftarrow R(s, a) + \gamma \max_{a' \in \mathcal{A}} \mathbb{E}\big[Q^{\psi}(s', a')\big] \tag{4}$$

In our work, we use Soft Actor-Critic (SAC) (Haarnoja et al., 2018) to learn Q^{ψ}. However, this could be substituted for any standard actor-critic or Q-learning approach. We note that if a biomechanical simulator is not available for a given task, the Q-function can be equivalently learned via offline RL on a dataset of human biomechanical rollouts in the environment. Importantly, through either a simulator or offline data, our world-action value model learns from only the biomechanical action output of a brain without relying on access to neural activity of the brain itself.

4.2. Training Brain Model

Given the world policy defined in Equation (3), our objective is to next learn to transform coprocessor actions into world actions. Because of the heterogeneity and complexity of the human brain, this process cannot be easily simulated and must instead be learned online. We utilize the Q-function learned in the previous step to guide online learning (Algorithm 1 line 7). We aim to select stimulations $\bar{a}$ that produce world actions of maximum value, thereby focusing learning of $\widehat{F}^{\phi}_{\text{brain}}$ on high-value regions of $\mathcal{A}$. After each patient interaction, we collect an experience, $(s, \bar{a}, r, s')$, which we use to update $\widehat{F}^{\phi}_{\text{brain}}$. Our sampling strategy is defined in Equation (5). After each interaction, we retrain $\widehat{F}^{\phi}_{\text{brain}}$ based upon our collected set of experiences (Algorithm 1 line 9).

$$\bar{\pi}(s) \triangleq \arg\max_{\bar{a} \in \bar{\mathcal{A}}} \mathbb{E}\big[Q^{\psi}\big(s, \widehat{F}^{\phi}_{\text{brain}}(s, \bar{a})\big)\big] \tag{5}$$

We update the brain model via the mean-squared error loss between the predicted world action, $\hat{a}$ and the ground truth world action a. To effectively capture the complexity of the relationship between stimulations and world actions, we adopt a structure for $\widehat{F}^{\phi}_{\text{brain}}(s, \bar{a})$ akin to the model presented (Bryan et al., 2022). In Bryan et al. (2022), the authors leverage a neural network to learn the mapping from stimulations to world actions from a monkey stroke dataset and show that this model is able to effectively capture the effects of stimulation on the brain

4.3. Updating world-action value model

The remaining issue to correct is that Equation (5) maximizes Q^{ψ} under the model $\widehat{F}^{\phi}_{\text{brain}}$. Unfortunately, $\widehat{F}^{\phi}_{\text{brain}}$ will not be a perfect model of the effects of stimulation and even if it were, not all actions $a \in \mathcal{A}$ are necessarily realizable by

stimulation $\bar{a}$. Thus, Q^{ψ} will be myopically over-optimistic when predicting Q-values from the perspective of the coprocessor agent. To solve this problem, we must continuously recalibrate the Q-function based on which actions can be realized by stimulation. We perform this calibration in the simulation MDP $\widehat{M}$ using the update in Equation (6). This procedure is illustrated in Algorithm 1, lines 10-14.

$$Q^{\psi}(s,a) \leftarrow r + \gamma \max_{\bar{a} \in \bar{\mathcal{A}}} \mathbb{E}\left[Q^{\psi}(s', \widehat{F}^{\phi}_{\text{brain}}(s', \bar{a}))\right] \quad (6)$$

In summary, we first train the world-action value model Q^{ψ} offline, and then iteratively update it while also learning $\widehat{F}^{\phi}_{\text{brain}}$. Repeating the last two stages (training brain model and updating world-action value model) enables us to learn an effective coprocessor policy via minimal interactions with the patient. We call our approach Coprocessor Actor Critic (CopAC) and demonstrate its performance in comparison with other RL methods in the next section.

5. Experiments

To aid a patient in accomplishing a task such as reaching and grasping an object, a coprocessor must learn a patient-specific stimulation policy in both an efficient and effective manner. Thus, the goal of our experimental evaluation is to 1) analyze the sample efficiency of CopAC compared to state-of-the-art MFRL and MBRL baselines and 2) investigate the reward attained by CopAC in comparison to these baselines.

We compare CopAC to the popular MFRL approach, Soft Actor-Critic (SAC) (Haarnoja et al., 2018), which combines actor and critic networks with an entropy regularization term, promoting exploration in a stable and efficient manner. We choose to compare to SAC because actor-critic algorithms have been employed in prior work in learning a policy for closed-loop brain stimulation (Gao et al., 2023; Lu et al., 2020a). We additionally baseline against the MBRL approach Model-Based Policy Optimization (Janner et al., 2021), which trains a model of the environment and uses both real experience and simulated experience from the model to update its policy. To assess the effectiveness of our sampling policy and the importance of updating the world-action value model, we conduct an ablation experiment and compare against CopAC with a random sampling policy (instead of maximizing the Q-function) as well as CopAC without updating the world-action value model.

5.1. *In Silico* Evaluation Environments

Evaluating novel RL approaches for adaptive brain stimulation *in vivo* is risky for patients and may waste patients' valuable time if the approach is not successful. Thus, it is common practice to first conduct experiments *in silico* (i.e., in simulation) to verify the efficacy of the approach before deploying in patients (Little et al., 2013a; Ashmaig et al., 2018). In our experiments, we investigate the ability of our method to restore the functionality of a synthetic injured brain across a range of simulated control tasks. The goal in each environment is to learn a coprocessor control policy to provide the appropriate stimulation to the brain to recover function and improve performance on the tasks post-injury. Below we discuss the control tasks and the synthetic brain models employed in our *in silico* experiments.

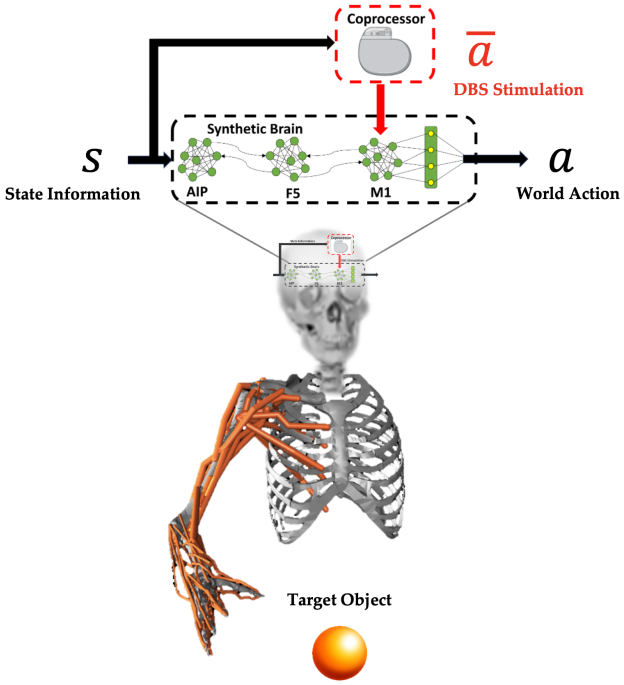

Figure 2. This figure shows the brain stimulation domain for the MyoSim Arm Reach task. We model the biomechanics of the reaching tasks using the MyoSuite physics simulator (Caggiano et al., 2022). The brain of a stroke patient is modeled via the approach described by Michaels et al. (2020) and consists of the anterior intraparietal area (AIP), ventral premotor cortex (F5), and primary motor cortex (M1) modules. The coprocessor applies stimulation to the motor cortex (M1) which modifies the world action of the patient.

Physiologically and Neurologically Realistic Stroke Domain: Drawing on prior work in neurophysiological and biomechanical modeling (Michaels et al., 2020; Caggiano et al., 2022), we introduce a novel simulation domain for evaluating adaptive brain stimulation policies in stroke patients (Figure 2). Such an *in silico* evaluation requires both a neurologically realistic human brain model of a stroke patient and a high-fidelity biomechanical simulator. To simulate the biomechanics of the human musculoskeletal system, we rely on MyoSuite (Caggiano et al., 2022), a cutting-edge simulator for biomechanical control problems based on the MuJoCo physics engine. The action space con-

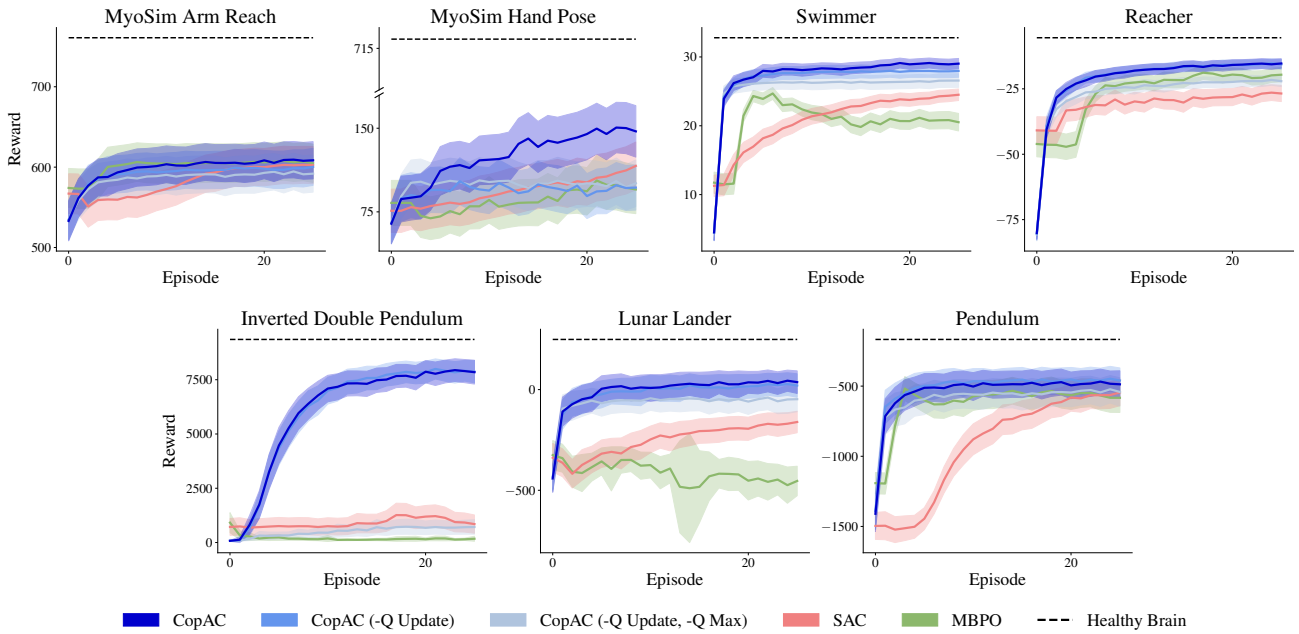

Figure 3. Evaluation results for CopAC compared to SAC, MBPO, and ablated CopAC. The dashed line represents the reward obtained by the healthy brain.

sists of individual muscle activations and the observation space consists of joint angles.

We simulate the injured brain of a stroke patient using the cortical brain model developed by Michaels et al. (2020). The backbone of this model is a recurrent neural network that mimics the modular structure of the anatomical circuit encompassing the visual cortex, the anterior intraparietal area (AIP), the ventral premotor cortex (F5), and the primary motor cortex (M1). Michaels et al. (2020) show that the emergent neural dynamics of this model correspond to the neural responses exhibited in a non-human primate's brain. Our insight is that we can employ standard RL techniques to train this brain model to accomplish various physical tasks in the MyoSuite biomechanical simulation, thus establishing a realistic pipeline between brain motor control signals, muscle kinematics, and biomechanical movement. Michaels et al. (2020) additionally demonstrate that zeroing a portion of the weights in the desired brain structure can reproduce behavioral deficits caused by stroke-induced brain lesions. By lesioning the simulated brain following the protocol detailed by Michaels et al. (2020), we induce stroke-like neurological and physiological behavior.

This method can be employed to simulate stroke patient control strategies across various functional benchmark tasks (e.g., in-hand object manipulation, object grasping, ambulation, visual-spatial acuity, etc.). We choose to evaluate CopAC on a reaching task that requires goal-directed functional movement (Figure 2). Such spatial reaching tasks, clinically known as task-related reaching training, are common benchmark tasks in which stroke patients often exhibit suboptimal control strategies (Thielman et al., 2004). We also evaluate CopAC on a spatial pose task in which the goal is to move the fingers to target locations. This task requires fine motor skills and emulates activities such as grasping an object that can be challenging for stroke patients to execute. (Buhmann et al., 2017).

Given a neurophysiologically realistic model of a stroke patient, the last step is to simulate the effects of closed-loop stimulation on the lesioned brain. Bryan et al. (2022) introduce a method to spatially and temporally simulate brain stimulation in the primary motor cortex to approximate the effect of *in vivo* stimulation. We rely on this approach to simulate the coprocessor's effect on a stroke patient's brain.

Standard Continuous Control Tasks: We additionally evaluate CopAC on a variety of standard continuous control tasks from the OpenAI Gym benchmark suite (Brockman et al., 2016). Although the nature of these tasks is distinct from the intricate control of human movement that is typical of adaptive brain stimulation tasks, our objective in scrutinizing our approach within these domains is twofold: firstly, to showcase its adaptability in handling a variety of complex tasks with varying state and action space dimensions, and secondly, to provide a benchmark comparison against previous approaches in well-established domains.

For OpenAI Gym benchmark tasks, we simulate a control policy generated by the brain of a stroke patient by first training a neural network policy using standard RL techniques to solve each Gym environment. We then "injure"

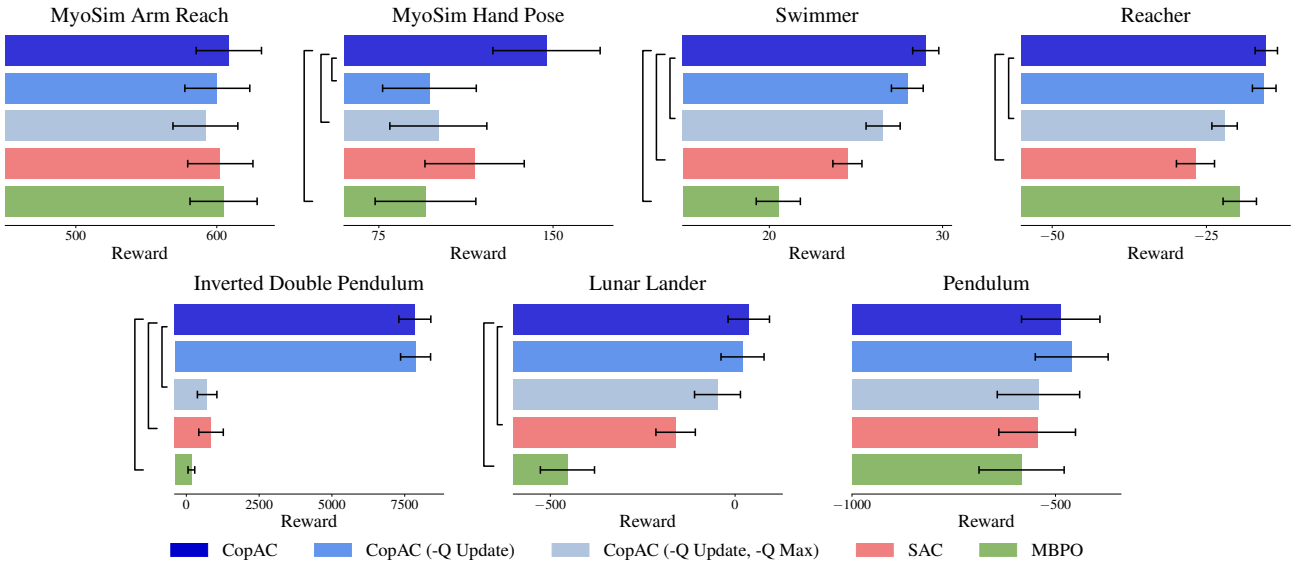

Figure 4. Evaluation results for CopAC compared to SAC and MBPO, and ablated CopAC. We display the evaluation reward after 25 episodes of training. Statistically significant differences between CopAC and other methods are marked with brackets.

the policy by zeroing random weights between two hidden layers in the network. Coprocessor stimulation is applied by additively modifying the values of randomly selected neurons in the layer following lesion.

5.2. Evaluation Reward

Figure 3 shows a comparison of the evaluation reward after each episode of online training for each of the approaches. The dashed line represents the performance of an unaffected or "healthy" brain, without any lesion. We note that we do not necessarily expect a stimulation policy to be capable of achieving healthy performance, as the severity of the lesion may limit the brain's ability to reach healthy-level performance. Instead, the primary objective is to provide stimulation to assist the patient in attaining a reward as close to a healthy level as possible.

We average our results across a set of injuries ranging from 0% to 100% lesion of the motor cortex for 25 episodes of online interaction. We limit the number of episodes to 25 as requiring a stroke patient to perform a greater number of task repetitions would likely be too physically and mentally demanding. Despite this tight sampling constraint, CopAC is able to quickly learn an effective strategy in less than 25 episodes in each of the control environments. On the contrary, SAC often struggles to improve performance significantly beyond random sampling within the limited time frame. Even when SAC is able to achieve a strategy comparable to CopAC, it typically requires double the number of interactions to do so. We find that MBPO outperforms SAC in terms of evaluation reward. In environments in which the model is simpler to learn such as MyoSim Reach and Pendulum, MBPO performs on par with CopAC. However, when dealing with more complex environments such as MyoSim Pose in which each finger must be precisely manipulated, MBPO learns much slower than our approach.

We next investigate the importance of the various components of CopAC via an ablation study. We compare the evaluation rewards of CopAC with two modified versions: CopAC (-Q Update) and CopAC (-Q Update, -Q Max). CopAC (-Q Update) does not update the world-action value model during online learning. We anticipate that not including this update will only impact performance when not all world actions are realizable due to brain lesioning. We see the most prominent benefit of updating the world-action model in the Swimmer domain. We also see an improvement in Reacher and a small improvement towards the end of the 25 episodes in several other domains.

We next investigate the benefit of our sampling strategy which selects the stimulation that maximizes Q^{ψ}, given our current knowledge of $\widehat{F}^{\phi}_{\text{brain}}$. To do so, we compare CopAC to a random sampling strategy without updating the world action-value model, i.e., CopAC (-Q update, -Q max). Because we are not focusing sampling on high-value regions of the world actions space, we expect CopAC (-Q update, -Q max) to learn more slowly compared to CopAC. We find that a random sampling strategy under-performs in the MyoSim Pose domain compared to CopAC's strategic sampling strategy and also produces lower evaluation reward in the Swimmer, Reacher, Pendulum, and Inverted Double Pendulum domains.

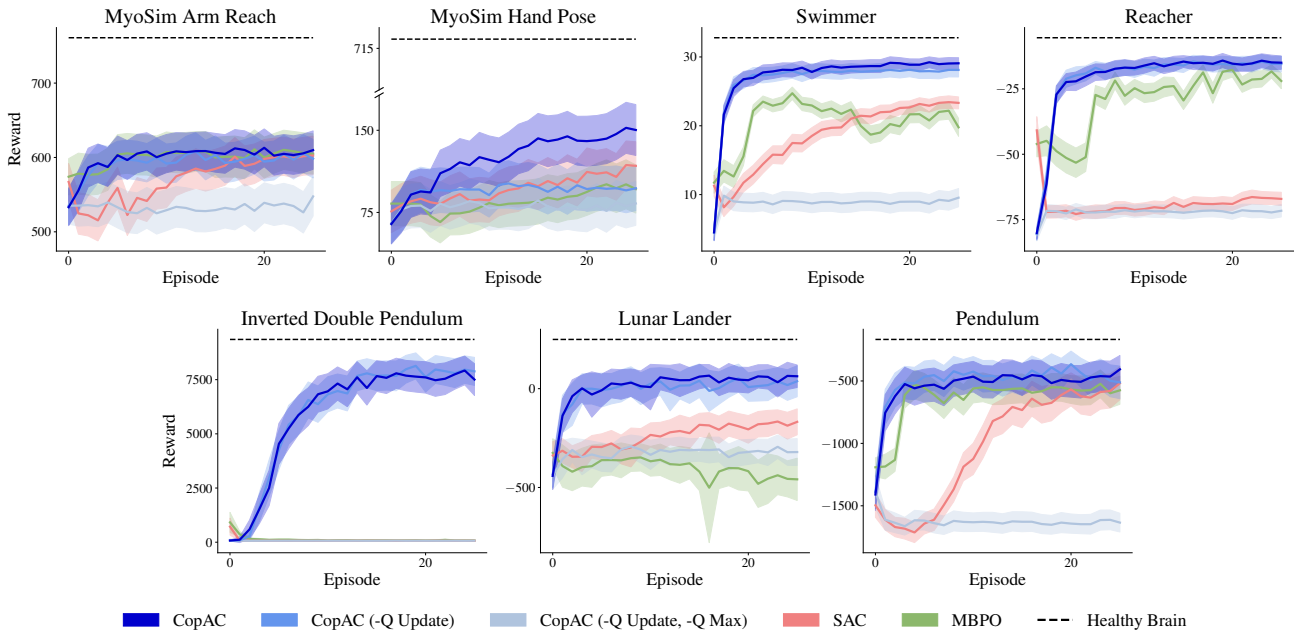

Figure 5. Training results for CopAC compared to SAC, MBPO, and ablated CopAC. The dashed line represents the reward obtained by the healthy brain.

5.3. Training Reward

Ideally, a coprocessor should simultaneously provide stimulation that maximally assists the stroke patient given its current model of the patient while continuously refining and updating its policy. In the previous section, we established that our approach learns a superior policy with fewer samples than the baseline methods. Here we investigate if, during policy learning, we are able to achieve a higher training reward compared to baselines. Training reward is an important metric in our brain stimulation domain because it indicates how well an approach is able to aid a patient in completing the desired task during policy learning. Figure 5 shows the training reward summed over each episode of online interaction. Due to our sampling strategy which selects stimulations that produce high-value world actions, our approach is able to achieve high task reward during learning. This means that CopAC can successfully assist the patient in accomplishing a given task during the learning phase. In contrast, MBPO and SAC achieve a significantly lower training reward. CopAC (-Q update, -Q max) performs poorly in all environments due to the random sampling policy.

6. Discussion

Our findings demonstrate the advantage of MBRL over MFRL in the domain of adaptive brain stimulation. Leveraging a model of the patient's biomechanics enables CopAC to reduce the number of interactions and achieve more than a 10x benefit in sample efficiency in many environments. This improved sample efficiency means that patients will be able to quickly benefit from a high-quality stimulation policy. We find that the benefit of our approach is particularly apparent in domains that require complex motor movements with large action spaces (e.g., MyoSim Hand Pose) whereas domains that require more coarse movements with lower dimensional action spaces (e.g., MyoSim Arm Reach) do not produce as large of a benefit compared to baselines. This insight is important because stroke patients often struggle with complex tasks requiring fine motor skills. We show that our approach is the most effective at learning stimulation policies for these complex tasks whereas an approach like MBPO may suffice for simpler tasks that require only gross motor control.

In the adaptive brain stimulation domain, high training reward is important for improving patient experience. A sampling strategy that produces off-task behavior will likely be disruptive and frustrating to a patient and could even pose a risk to the patient's safety. In our ablation study, we show that by leveraging the action-value model to sample in an on-policy manner, CopAC produces higher training reward compared to random sampling. CopAC's training reward contrasts with both SAC and MBRL which utilize an off-policy strategy to sample stimulations, resulting in lower reward. Notably, CopAC's strategy exhibits a significant advantage in the MyoSim Hand Pose environment. We hypothesize that this result is due to the fact that this environment has a much larger action space (39-dimensional) compared to the other environments. This result underscores the importance of a strategic sampling approach to

efficiently learn a high-quality policy in environments with a substantial action space.

We find that updating the world action-value model is an important component in some domains. However, if all world actions are realizable by the injured brain or if the unrealizable actions are unimportant to task success, then updating the world action-value model may be inconsequential.

Our results show strengths of CopAC in areas particularly crucial for adaptive brain stimulation. The learned world action-value model enables good performance throughout training, which is important for patient comfort and safety. Our approach additionally excels over baselines in complex tasks requiring fine-grained control, providing benefits in both sample efficiency and performance.

7. Conclusion

Summary: We have introduced CopAC, an MBRL approach for learning a coprocessor policy for stroke patients. Our key insight is that the optimal stimulation policy is a combination of modeling optimal world actions and determining how to produce world actions via brain stimulation. Our approach leverages a simulation-derived Q-function to model the quality of world actions for a given task. We then employ this world action-value model to intelligently and efficiently learn the mapping from coprocessor stimulations to world actions. To avoid an overly-optimistic Q-function, we iteratively update the action-value model based upon our current model of the brain. We demonstrate our approach in both a novel, physiologically-motivated environment and standard control tasks. Our approach excels in terms of sample efficiency and overall task reward, surpassing both MBRL and MFRL methods across all domains. By improving sample efficiency 10-fold, we take a step towards an RL approach that can be deployed *in vivo*. Our hope is to pave the way for future advancements in applying RL to closed-loop brain stimulation in real-world settings.

Limitations: One limitation of our work is that we only evaluate CopAC *in silico* due to the difficulty of *in vivo* evaluations. The invasive nature of brain stimulation and potential negative side effects of unsafe stimulation motivate us to first validate the efficacy of our approach in simulation before deploying in the real world. We are also limited by our reliance on a realistic biomechanical simulator to enable sample efficiency. Since we test in simulation, we have access to the same environment to learn an action-value model in the first step of CopAC. Our experiments therefore do not perfectly reflect real-world challenges posed by the gap in realism between simulators and human patients. Another challenge of real-world applications is the technology required (e.g., electromyography and vicon) for estimating world states and actions. Finally, although our approach minimizes patient interaction, it still requires online learning for which we have not theoretically guaranteed safety. Such safety guarantees would be crucial for any *in vivo* applications.

Future Work: Future work will focus on addressing these limitations. We will test how the action-value model trained in biomechanical simulation transfers to human patients in *in vivo* evaluations. Additionally, we will explore offline RL for learning the action-value model in circumstances where a biomechanical simulator is not available. To improve safety and mitigate patient risk during online learning, in future work we aim to draw upon existing approaches in safe RL and we will work closely with clinical collaborators to ensure that we are safely and appropriately constraining the stimulation space during *in vivo* experiments (García & Fernández, 2015).

Impact Statement

A major ethical concern of automated coprocessor policy learning is that adaptive brain stimulation can pose a safety risk to patients if stimulations outside of a safe region are applied to the brain. To mitigate these risks, in future work, when deploying CopAC on real patients, we aim to work closely with clinical collaborators to ensure patient safety and appropriately constrain the action space during online learning.

Another ethical consideration involves the possibility of RL coprocessor approaches being exploited for malicious control over end-users. Unauthorized manipulation of RL policies, such as through adversarial attacks, could lead to unethical interventions and compromise the well-being of individuals with brain implants. To reduce this risk, in future work, we will draw on prior work in guarding against adversarial attacks to mitigate potential exploitation (Chen et al., 2019).

Acknowledgements

This work was supported in part by a grant from the Weill Neurohub and by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

We thank Ian Heimbuch and Matthew Bryan for early discussions on this work.

References

Abbeel, P., Quigley, M., and Ng, A. Y. Using inaccurate models in reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 1–8, New York, NY, USA, 2006.

Association for Computing Machinery. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143845. URL https://doi.org/10.1145/1143844.1143845.

Ashmaig, O., Connolly, M., Gross, R. E., and Mahmoudi, B. Bayesian Optimization of Asynchronous Distributed Microelectrode Theta Stimulation and Spatial Memory. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2018-July:2683–2686, 2018. ISSN 1557170X. doi: 10.1109/EMBC.2018.8512801.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Bronte-Stewart, H. M., Petrucci, M. N., O'Day, J. J., Afzal, M. F., Parker, J. E., Kehnemouyi, Y. M., Wilkins, K. B., Orthlieb, G. C., and Hoffman, S. L. Perspective: Evolution of control variables and policies for closed-loop deep brain stimulation for parkinson's disease using bidirectional deep-brain-computer interfaces. *Frontiers in Human Neuroscience*, 14, 8 2020. ISSN 16625161. doi: 10.3389/fnhum.2020.00353.

Bryan, M. J., Jiang, L. P., and Rao, R. P. N. Neural coprocessors for restoring brain function: Results from a cortical model of grasping. *Journal of Neural Engineering*, 10 2022. URL http://arxiv.org/abs/2210.11478.

Buhmann, C., Huckhagel, T., Engel, K., Gulberti, A., Hidding, U., Poetter-Nerger, M., Goerendt, I., Ludewig, P., Braass, H., Choe, C. U., Krajewski, K., Oehlwein, C., Mittmann, K., Engel, A. K., Gerloff, C., Westphal, M., Köppen, J. A., Moll, C. K. E., and Hamel, W. Adverse events in deep brain stimulation: A retrospective long-term analysis of neurological, psychiatric and other occurrences. *PLoS ONE*, 12, 7 2017. ISSN 19326203. doi: 10.1371/journal.pone.0178984.

Caggiano, V., Wang, H., Durandau, G., Sartori, M., and Kumar, V. MyoSuite – a contact-rich simulation suite for musculoskeletal motor control, 2022.

Chen, T., Liu, J., Xiang, Y., Niu, W., Tong, E., and Han, Z. Adversarial attack and defense in reinforcement learning-from AI security view. *Cybersecurity*, 2, 12 2019. ISSN 25233246. doi: 10.1186/s42400-019-0027-x.

Choi, H., Park, D., Rha, D. W., Nam, H. S., Jo, Y. J., and Kim, D. Y. Kinematic analysis of movement patterns during a reach-and-grasp task in stroke patients. *Frontiers in Neurology*, 14, 2023. ISSN 16642295. doi: 10.3389/fneur.2023.1225425.

Coventry, B. S. and Bartlett, E. L. Closed-loop reinforcement learning based deep brain stimulation using SpikerNet: A computational model. In *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 1–4, 2023. doi: 10.1109/NER52421.2023.10123797.

Cunha, C. D., Boschen, S. L., Gómez-A, A., Ross, E. K., Gibson, W. S. J., Min, H.-K., Lee, K. H., and Blaha, C. D. Toward sophisticated basal ganglia neuromodulation: Review on basal ganglia deep brain stimulation. *Neuroscience & Biobehavioral Reviews*, 58:186–210, November 2015. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2015.02.003.

Delp, S. L., Anderson, F. C., Arnold, A. S., Loan, P., Habib, A., John, C. T., Guendelman, E., and Thelen, D. G. OpenSim: Open-source software to create and analyze dynamic simulations of movement. *IEEE Transactions on Biomedical Engineering*, 54(11):1940–1950, 2007. doi: 10.1109/TBME.2007.901024.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110:2419–2468, 9 2021. ISSN 15730565. doi: 10.1007/s10994-021-05961-4.

Elias, G. J. B., Namasivayam, A. A., and Lozano, A. M. Deep brain stimulation for stroke: Current uses and future directions. *Brain Stimulation*, 11(1):3–28, 2018. ISSN 1935-861X. doi: https://doi.org/10.1016/j.brs.2017.10.005. URL https://www.sciencedirect.com/science/article/pii/S1935861X17309361.

Frey, J., Cagle, J., Johnson, K. A., Wong, J. K., Hilliard, J. D., Butson, C. R., Okun, M. S., and de Hemptinne, C. Past, present, and future of deep brain stimulation: Hardware, software, imaging, physiology and novel approaches. *Frontiers in Neurology*, 13, 3 2022. ISSN 16642295. doi: 10.3389/fneur.2022.825178.

Ganguly, K., Khanna, P., Morecraft, R. J., and Lin, D. J. Modulation of neural co-firing to enhance network transmission and improve motor function after stroke. *Neuron*, 110(15):2363–2385, August 2022. ISSN 08966273. doi: 10.1016/j.neuron.2022.06.024.

Gao, Q., Schmidt, S. L., Chowdhury, A., Feng, G., Peters, J. J., Genty, K., Grill, W. M., Turner, D. A., and Pajic, M. Offline learning of closed-loop deep brain stimulation controllers for parkinson disease treatment. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, ICCPS '23, pp. 44–55, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 979-8400-70-0-3-6-1. doi: 10.1145/3576841.3585925. URL https://doi.org/10.1145/3576841.3585925.

García, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.*, 16:1437–1480, 2015. URL https://api.semanticscholar.org/CorpusID:2497153.

Ghasemi, P., Sahraee, T., and Mohammadi, A. Closed-and open-loop deep brain stimulation: Methods, challenges, current and future aspects. *Journal of Biomedical Physics and Engineering*, 8:209–216, 2018. URL www.jbpe.org.

Groiss, S. J., Wojtecki, L., Sudmeyer, M., and Schnitzler, A. Deep brain stimulation in parkinson-s disease. *Therapeutic Advances in Neurological Disorders*, 2:379–391, 2009. ISSN 17562856. doi: 10.1177/1756285609339382.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018.

Hatem, S. M., Saussez, G., della Faille, M., Prist, V., Zhang, X., Dispa, D., and Bleyenheuft, Y. Rehabilitation of motor function after stroke: A multiple systematic review focused on techniques to stimulate upper extremity recovery. *Frontiers in Human Neuroscience*, 10, 9 2016. ISSN 16625161. doi: 10.3389/fnhum.2016.00442.

Hu, W. and Stead, M. Deep brain stimulation for dystonia. *Translational Neurodegeneration*, 2014. URL http://www.translationalneurodegeneration.com/content/3/1/2.

Huang, Q. Model-based or model-free, a review of approaches in reinforcement learning. In *2020 International Conference on Computing and Data Science (CDS)*, pp. 219–221, 2020. doi: 10.1109/CDS49703.2020.00051.

Ingwersen, T., Wolf, S., Birke, G., Schlemm, E., Bartling, C., Bender, G., Meyer, A., Nolte, A., Ottes, K., Pade, O., Peller, M., Steinmetz, J., Gerloff, C., and Thomalla, G. Long-term recovery of upper limb motor function and self-reported health: results from a multicenter observational study 1 year after discharge from rehabilitation. *Neurological Research and Practice*, 3, 12 2021. ISSN 25243489. doi: 10.1186/s42466-021-00164-7.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization, 2021.

Khanna, P., Totten, D., Novik, L., Roberts, J., Morecraft, R. J., and Ganguly, K. Low-frequency stimulation enhances ensemble co-firing and dexterity after stroke. *Cell*, 184(4):912–930.e20, February 2021. ISSN 00928674. doi: 10.1016/j.cell.2021.01.023.

Little, S., Pogosyan, A., Neal, S., Zavala, B., Zrinzo, L., Hariz, M., Foltynie, T., Limousin, P., Ashkan, K., FitzGerald, J., Green, A. L., Aziz, T. Z., and Brown, P. Adaptive deep brain stimulation in advanced parkinson disease. *Annals of neurology*, 74:449–457, 2013a. ISSN 15318249. doi: 10.1002/ana.23951.

Little, S., Pogosyan, A., Neal, S., Zavala, B., Zrinzo, L., Hariz, M., Foltynie, T., Limousin, P., Ashkan, K., FitzGerald, J., et al. Adaptive deep brain stimulation in advanced Parkinson disease. *Annals of Neurology*, 74(3): 449–457, September 2013b. ISSN 1531-8249. doi: 10.1002/ana.23951.

Lozano, A. M., Lipsman, N., Bergman, H., Brown, P., Chabardes, S., Chang, J. W., Matthews, K., McIntyre, C. C., Schlaepfer, T. E., Schulder, M., Temel, Y., Volkmann, J., and Krauss, J. K. Deep brain stimulation: current challenges and future directions. *Nature Reviews Neurology*, 15:148–160, 3 2019. ISSN 17594766. doi: 10.1038/s41582-018-0128-2.

Lu, M., Wei, X., Che, Y., Wang, J., and Loparo, K. A. Application of reinforcement learning to deep brain stimulation in a computational model of parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(1):339–349, 2020a. doi: 10.1109/TNSRE.2019.2952637.

Lu, M., Wei, X., Che, Y., Wang, J., and Loparo, K. A. Application of Reinforcement Learning to Deep Brain Stimulation in a Computational Model of Parkinson's Disease. *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, 28(1):339–349, January 2020b. ISSN 1558-0210. doi: 10.1109/TNSRE.2019.2952637.

Michaels, J. A., Schaffelhofer, S., Agudelo-Toro, A., and Scherberger, H. A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *The Proceedings of the National Academy of Sciences*, 2020. doi: 10.1073/pnas.2005087117/-/DCSupplemental. URL https://www.pnas.org.

Moerland, T. M., Broekens, J., Plaat, A., and Jonker, C. M. Model-based reinforcement learning: A survey, 2022.

Oehrn, C. R., Cernera, S., Hammer, L. H., Shcherbakova, M., Yao, J., Hahn, A., Wang, S., Ostrem, J. L., Little, S., and Starr, P. A. Personalized chronic adaptive deep brain stimulation outperforms conventional stimulation in parkinson's disease. *medRxiv : the preprint server for health sciences*, 8 2023. doi: 10.1101/2023.08.03.23293450. URL http://www.ncbi.nlm.nih.gov/pubmed/37649907http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC10463549.

Rao, R. P. Towards neural co-processors for the brain: combining decoding and encoding in brain–computer interfaces. *Current Opinion in Neurobiology*, 55:142–151, 4 2019. ISSN 18736882. doi: 10.1016/j.conb.2019.03.008.

Schrum, M., Connolly, M. J., Cole, E., Ghetiya, M., Gross, R., and Gombolay, M. C. Meta-active learning in probabilistically safe optimization. *IEEE Robotics and Automation Letters*, 7(4):10713–10720, 2022.

Thielman, G. T., Dean, C. M., and Gentile, A. M. Rehabilitation of reaching after stroke: Task-related training versus progressive resistive exercise11No commercial party having a direct interest in the results of the research supporting this article has or will confer a benefit on the author(s) or on any organization with which the author(s) is/are associated. *Archives of Physical Medicine and Rehabilitation*, 85(10):1613–1618, 2004. ISSN 0003-9993. doi: https://doi.org/10.1016/j.apmr.2004.01.028. URL https://www.sciencedirect.com/science/article/pii/S0003999304002916.

Valencia, D., Jia, J., Li, R., Hayashi, A., Lecchi, M., Terezakis, R., Gee, T., Liarokapis, M., MacDonald, B. A., and Williams, H. Comparison of model-based and model-free reinforcement learning for real-world dexterous robotic manipulation tasks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 871–878, 2023. doi: 10.1109/ICRA48891.2023.10160983.

Visanji, N. P., Ghani, M., Yu, E., Kakhki, E. G., Sato, C., Moreno, D., Naranian, T., Poon, Y. Y., Abdollahi, M., Naghibzadeh, M., Rajalingam, R., Lozano, A. M., Kalia, S. K., Hodaie, M., Cohn, M., Statucka, M., Boutet, A., Elias, G. J. B., Germann, J., Munhoz, R., Lang, A. E., Gan-Or, Z., Rogaeva, E., and Fasano, A. Axial impairment following deep brain stimulation in parkinson's disease: A surgicogenomic approach. *Journal of Parkinson's Disease*, 12:117–128, 2022. ISSN 1877718X. doi: 10.3233/JPD-212730.

Appendix

A. Learning Q^ψ with offline RL

Our method hinges on learning Q^ψ without requiring online patient interactions. While we demonstrate that we can leverage a biomechanical simulator to learn Q^ψ, in some instances we may not have access to a high-fidelity simulator. In these cases, we consider the use of an existing dataset to train Q^ψ through offline RL. This approach assumes that we have access to a historical coprocessor dataset from stimulation policies previously deployed on the patient. We use Conservative Soft Actor Critic to learn Q^ψ from this dataset. Once Q^ψ is learned from the offline data, we follow the same procedure for learning $\widehat{F}^\phi_{\text{brain}}$ as discussed in Section 4.2.

Figure 6 shows the training reward for our approach when Q^ψ is learned via offline RL compared to baselines. We show that CopAC (offline) performs better than the baselines in most environments but performs slightly worse than CopAC when Q^ψ is learned via simulation. This outcome supports the viability of offline RL as an alternative approach. However, it suggests that using a biomechanical simulator, when available, is likely a better option for learning Q^ψ. In future work we aim to investigate how the amount of data and the suboptimality of the policy used to collect the data affects performance.

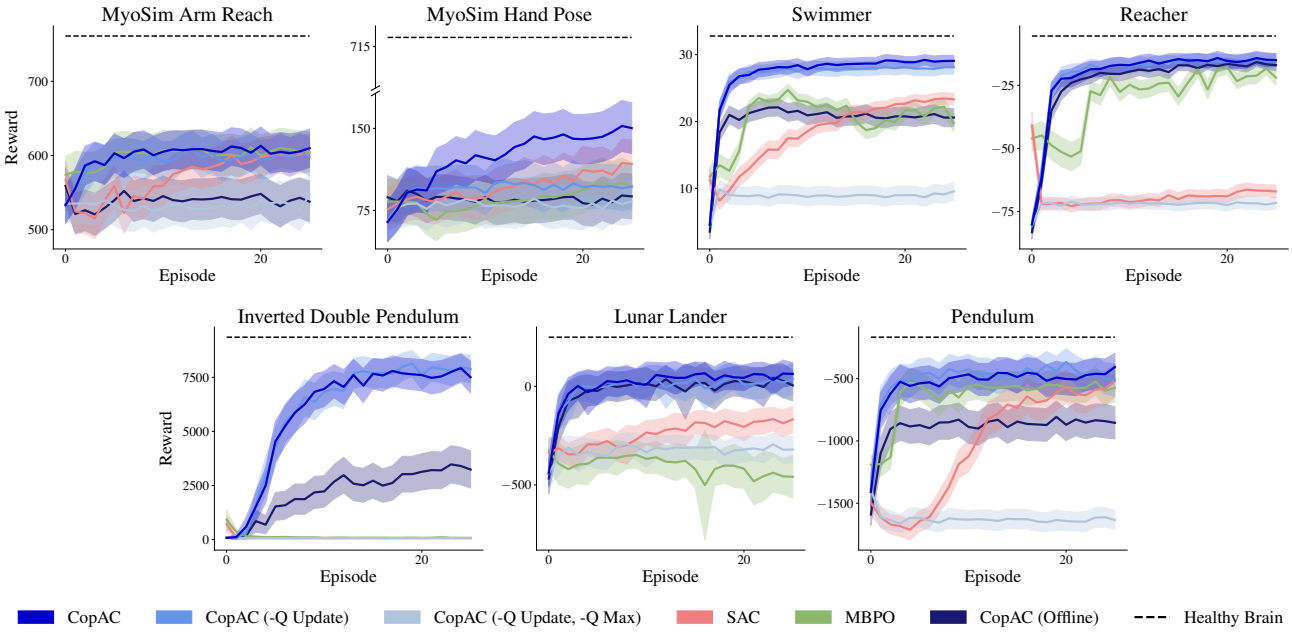

Figure 6. Training reward for CopAC compared to SAC, MBPO, and ablated CopAC. The dashed line represents the reward obtained by the healthy brain.

B. $\widehat{F}^\phi_{\text{brain}}$ training details

We train $\widehat{F}^\phi_{\text{brain}}$ for 75 epochs with a learning rate of $5 \cdot 10^{-3}$. The architecture consists of three hidden layers with ReLU activations consisting of 64, 32, and 8 neurons.

C. Robustness to initialization of $\widehat{F}^\phi_{\text{brain}}$

To validate that CopAC is robust to the initialization of $\widehat{F}^\phi_{\text{brain}}$, we additionally run CopAC and ablations across 5 random seeds. Results are displayed in Figures 7 and 8. Here, we only use a single brain for each environment rather than taking the average across multiple brains as in our other experiments.

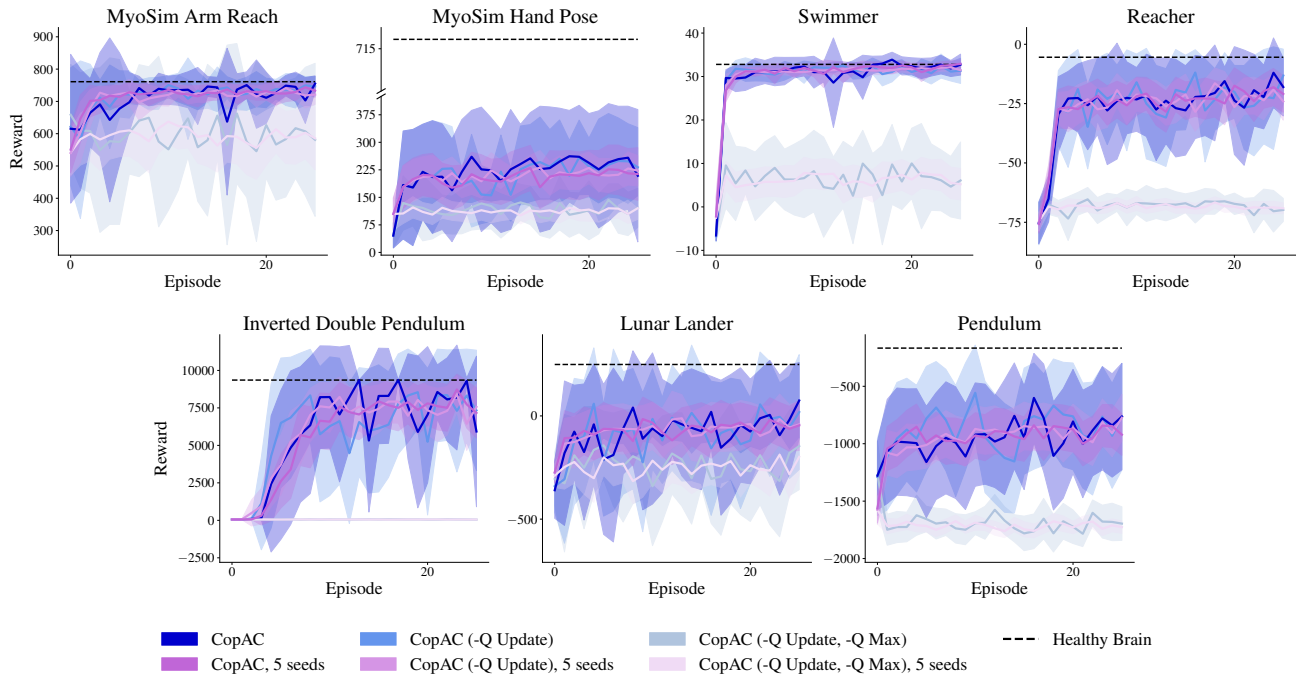

Figure 7. Training reward for CopAC and ablations. Results from a single seed are displayed alongside results averaged across random seeds.

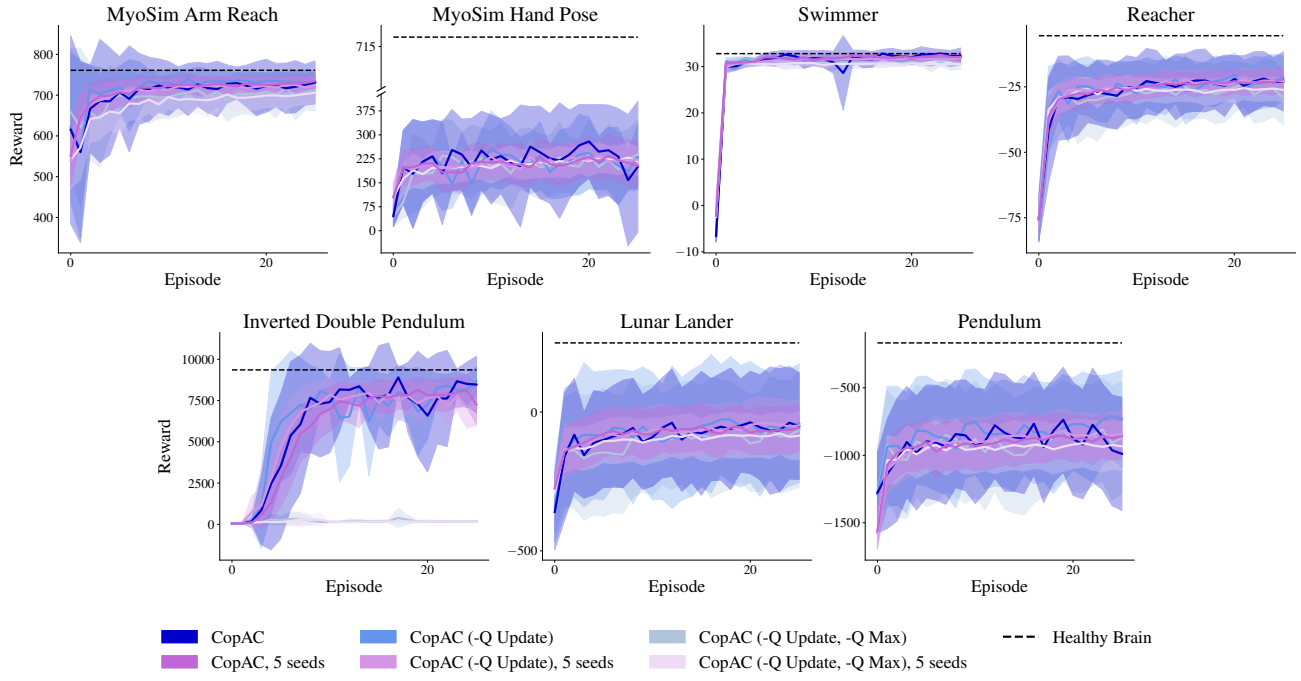

Figure 8. Evaluation reward for CopAC and ablations. Results from a single seed are displayed alongside results averaged across random seeds.

D. Comparison with inverse brain model coprocessor

We additionally compare CopAC to a baseline approach using an inverse brain model to select stimulations. We learn the inverse brain model $\widehat{F}^{\phi}_{\text{inverse}} : S \times \mathcal{A} \to \bar{\mathcal{A}}$ that maps world actions to the stimulations that would have induced them. The

inverse brain model coprocessor policy selects stimulations with $\widehat{F}^{\phi}_{\text{inverse}}$ to induce world actions returned by the optimal world policy π, as defined in Equation (7).

$$\bar{\pi}_{\text{inverse}}(s) \triangleq \widehat{F}^{\phi}_{\text{inverse}}(s, \pi(s)) \tag{7}$$

Algorithm 2 Inverse Brain Model Coprocessor

Require: world MDP $M = (\mathcal{S}, \mathcal{A}, P, R)$, $s_0 \sim p_0$
Require: stimulation space $\bar{\mathcal{A}}$, injured brain F_{brain}
initialize $\widehat{F}^{\phi}_{\text{inverse}}$
while access to simulator **do**
  rollout π in M with experiences (s, a, r, s')
  fit $Q^{\psi}(s, a)$ with Equation (4)
end while
while access to injured brain **do**
  rollout $\bar{\pi}_{\text{inverse}}$ in $\overline{M}$ with experiences $(s, \bar{a}, r, s')$
  $a \leftarrow F_{\text{brain}}(s, \bar{a})$
  fit $\widehat{F}^{\phi}_{\text{inverse}}(s, a)$ to $\bar{a}$
end while
return $\bar{\pi}_{\text{inverse}}$

The inverse brain model approach is presented in Algorithm 2. We evaluate it against CopAC and show a comparison of their performance during training and evaluation in Figures 9 and 10. We find that CopAC is able to achieve a higher reward and better sample efficiency compared to the inverse brain model.

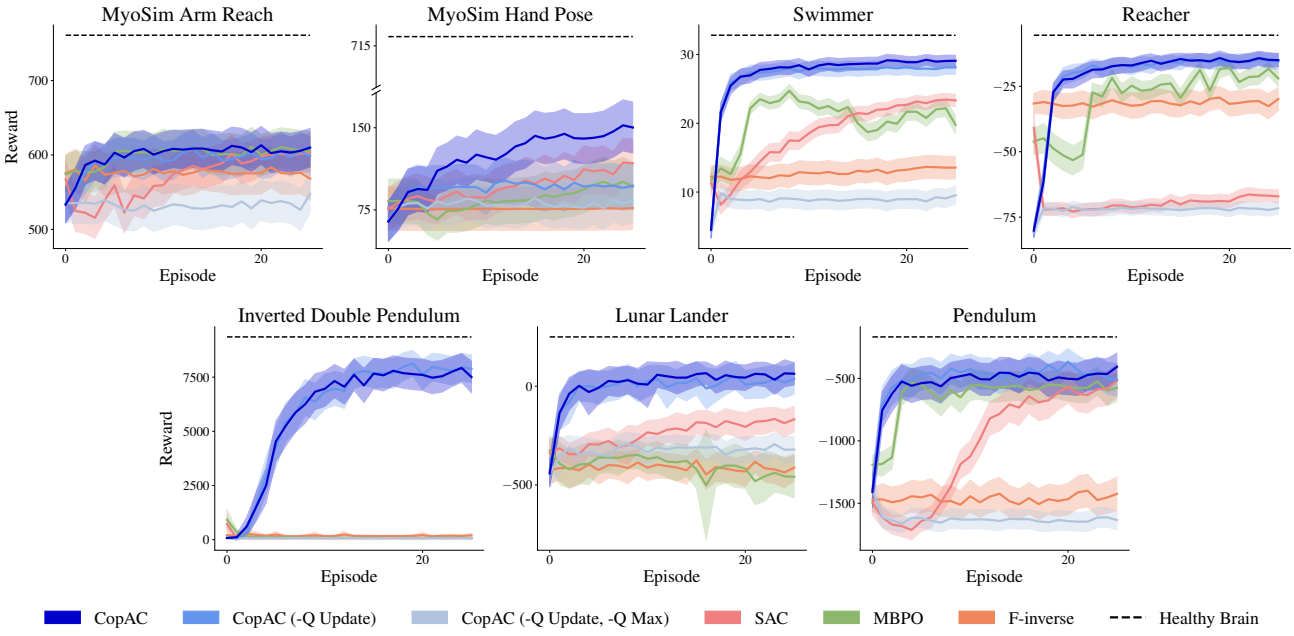

Figure 9. Training results for inverse brain model coprocessor compared to CopAC, ablated CopAC, and baselines.

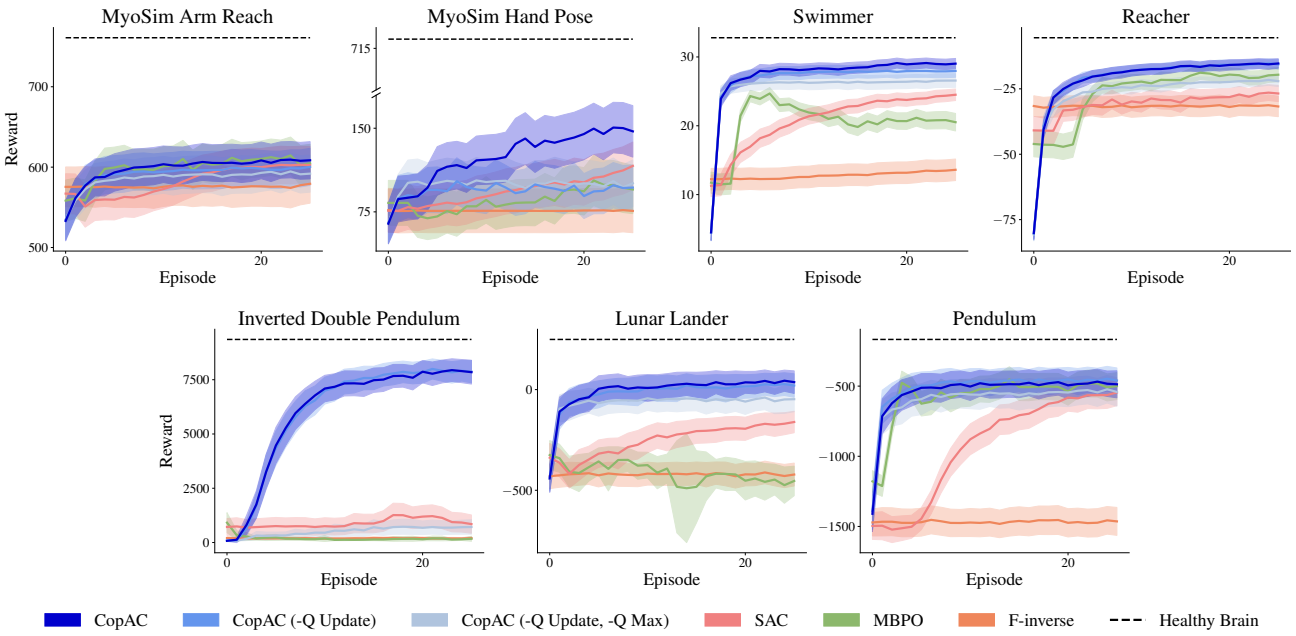

Figure 10. Evaluation results for inverse brain model coprocessor compared to CopAC, ablated CopAC, and baselines.

E. Robustness to sim to real gap

We perform experiments shown in Figure 11 to simulate the sim-to-real gap by systematically altering the dynamics of the environment during online training and testing while learning $\widehat{F}^{\phi}_{\text{brain}}$. We demonstrate that our method is robust to these effects up to a point and that the online training aspect of our approach can help to account for a mismatch between simulated and real biomechanics. Specifically, we show that we can alter gravity by up to 40% during online interaction in Pendulum and maintain performance. We show that our approach is robust to nearly 30% change in gravity in LunarLander. In our Myosim environments, we systematically alter the observations during online training and testing and show that the Arm Reach task maintains performance for 10% shift in observations and Hand Pose can handle nearly a 40% change.

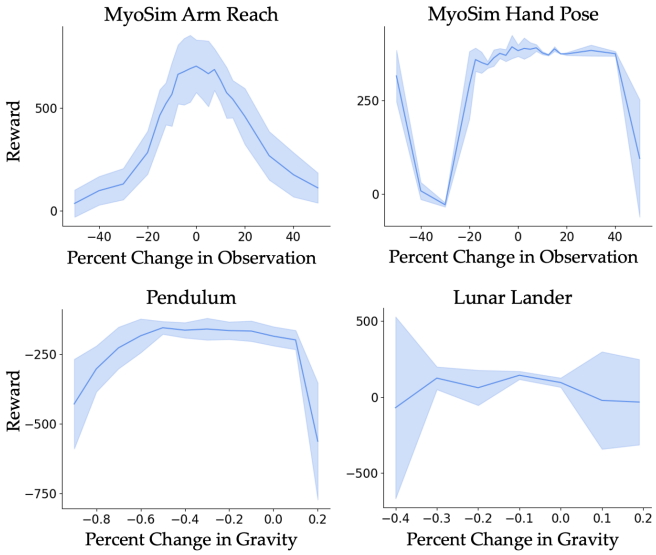

Figure 11. Evaluation reward when altering the environment during online training. We show that our approach is robust to a mismatch between simulation and reality.