Proceedings Track

Graph Neural Networks Uncover Geometric Neural Representations in Reinforcement-Based Motor Learning

Editors: List of editors' names

Abstract

Graph Neural Networks (GNN) can capture the geometric properties of neural representations in EEG data. Here we utilise those to study how reinforcement-based motor learning affects neural activity patterns during motor planning, leveraging the inherent graph structure of EEG channels to capture the spatial relationships in brain activity. By exploiting task-specific symmetries, we define different pretraining strategies that not only improve model performance across all participant groups but also validate the robustness of the geometric representations. Explainability analysis based on the graph structures reveals consistent group-specific neural signatures that persist across pretraining conditions, suggesting stable geometric structures in the neural representations associated with motor learning and feedback processing. These geometric patterns exhibit partial invariance to certain task space transformations, indicating symmetries that enable generalisation across conditions while maintaining specificity to individual learning strategies. This work demonstrates how GNNs can uncover the effects of previous outcomes on motor planning, in a complex real-world task, providing insights into the geometric principles governing neural representations. Our experimental design bridges the gap between controlled experiments and ecologically valid scenarios, offering new insights into the organisation of neural representations during naturalistic motor learning, which may open avenues for exploring fundamental principles governing brain activity in complex tasks.

Keywords: Motor Learning, Motor Planning, Neural Representations, Graph Neural Networks, EEG

1. Introduction

Motor learning, the acquisition and refinement of movement skills, is a fundamental aspect of human cognition and behaviour. Understanding its neural mechanisms is crucial for developing effective strategies in rehabilitation, sports training, and brain-computer interfaces. The use of EEG has been a longstanding approach in the investigation of neural activity underlying motor learning processes. EEG-based research has uncovered dynamic patterns of neural oscillations (Henz and Schöllhorn, 2016) and event-related synchronisation (Tan et al., 2014; Haar and Faisal, 2020) that correspond to the distinct learning mechanisms. These electrophysiological signatures have been associated with a variety of cognitive mechanisms essential to motor learning, such as attentional focus, sensorimotor integration, and memory consolidation (Talsma and Woldorff, 2005; Berti, 2008).

Reinforcement learning principles have provided insights into how the brain processes feedback and adapts behaviour based on previous experiences during motor learning. Neural activity patterns observed between trials have been found to reflect the processing of performance feedback from the previous trial, influencing subsequent motor planning and execution (Albert and Shadmehr, 2016; Luft, 2014; Yuan et al., 2015; Hou et al., 2022). This feedback-dependent modulation of brain activity underscores the iterative nature of motor skill acquisition and the importance of trial-to-trial learning in shaping neural representations (Eliassen et al., 2012).

© 2024 .

Related Work Traditional statistical analysis approaches were often limited in fully capturing the spatial connections between different brain regions, which are essential for understanding the distributed nature of neural processing during complex tasks (Zaepffel et al., 2013). However, recent advancements in machine learning have driven the development of more sophisticated techniques. Neural networks and deep learning have emerged as powerful tools for analysing EEG data, enabling the automatic extraction of complex spatial-temporal representations from high-dimensional EEG signals (Schirrmeister et al., 2017; Han et al., 2023a; Wang et al., 2018). Yet, these studies often face challenges in fully capturing the complex relationships between brain regions and ensuring the explainability of their findings.

Graph neural networks have emerged as a promising approach for analysing EEG data. These models can capture the temporal dynamics and spatial dependencies of neural activity, making them well-suited for studying the intricate network of brain connections involved in motor learning. However, the typical applications of deep learning have focused on static, well-structured brain data, such as that from motor imagery tasks (Hou et al., 2022; Han et al., 2023b; Sun et al., 2021). Applying these powerful models to dynamic real-world brain data remains an area to be explored. This has the potential to enhance our understanding of how the brain learns and adapts, informing the development of more effective training protocols, rehabilitation strategies, and brain-computer interfaces. Furthermore, insights gained from studying brain patterns of complex motor skills could have broader implications for our understanding of cognitive processes, strategy, decision-making, and adaptation relevant to various aspects of human behavior.

In this work, we present a novel application of GNNs to investigate the neural correlates of motor learning and planning in a real-world task. Based on an EEG dataset collected during a real-world billiard task in an Embodied Virtual Reality environment (Nardi et al., 2024), we trained a GNN model to predict the success or failure of the previous trial. The model leverages the EEG data from the pre-movement window, aiming to capture how the perceived success or failure of a given shot influences the neural processes involved in the motor planning of the next shot. By applying GNNs to EEG data collected during a dynamic real-world motor task, unlike previous studies which focused on static brain imaging and motor imagery, we captured the intricate neural dynamics associated with motor planning, execution, and feedback processing in a more naturalistic setting (Tsay et al., 2024). Furthermore, we used multiple explainability methods and conducted extensive analysis to identify which brain regions and functional interactions were most salient in the model's prediction. The results obtained contribute to our understanding of the neural mechanisms underlying motor skill acquisition and demonstrate the potential of GNNs as a powerful and explainable tool for the analysis of brain data.

2. Methods

Description of Dataset The dataset utilised in our research originates from experiments conducted by Nardi et al. (2024), where 40 healthy, inexperienced participants engaged in a real-world billiard task within an Embodied Virtual Reality (EVR) setup (Haar et al., 2021). During the task, the participants performed a pool shot with the objective of pocketing a target ball, while immersed in a virtual environment that was accurately aligned with a

physical setup, comprising a pool stick and a pool table. Throughout the task, brain activity was recorded with a wireless 14-channel EEG headset (Emotiv EPOC+) with a 256Hz sampling rate. The outcome variable examined was the *success* or *failure* during the previous trial, as we aimed to investigate any changes in brain activity resulting from the previous feedback.

The 2-second window preceding each shot was the focus of our analysis. This timeframe was selected to capture motor planning processes, occurring before the "main" movement, and to exclude any visual feedback from the model predictions, as participants observed the ball trajectory only after the shot was performed. Indeed, this window was chosen to avoid including signals unrelated to the primary movement, thereby directing our GNN model training to the most relevant data.

Participants took part in two learning rounds with different feedback conditions and shooting towards different pockets. Here we analyse brain activity only from the round with the error feedback task (Nardi et al., 2024), from four groups of 10 participants, differing by round and pocket: *First Left* (first round, shooting at the left pocket), *First Right* (first round, shooting at the right pocket), *Second Left* (second round, shooting towards the left pocket), and *Second Right* (second round, shooting towards the right pocket).

EEG Data Pre-processing EEG data was band-pass filtered at 1-45Hz using FIR filter, and cleaned by removing artifacts from the signals, including bad data segments and problematic channels, employing the *Artifact Subspace Reconstruction* algorithm (Kothe, 2013). Furthermore, signal windows were removed based on thresholds for the maximum standard deviation of bursts and the maximum fraction of contaminated channels allowed in the final output data for each considered window. The prefrontal channels (AF3, AF4) were removed from the EEG data to avoid the models wrongly classifying changes in eye movement patterns instead of the proper underlying brain signal, resulting in a 12-channel dataset. The data was segmented into 2-second epochs preceding each shot, as the motor planning window. An independent component analysis (ICA) was conducted on the data to eliminate muscle, eye, and cardiac artefacts. The missing channels were subsequently interpolated from the remaining ones using spherical spline interpolation and a topographical map. This preprocessing was informed by prior research on motor learning (e.g. Tan et al., 2014; Haar and Faisal, 2020) and based on the 10-20 international EEG system.

Due to the limited data available for each participant, we combined together the 10 participants within each group. We created a balanced dataset for each group by accounting for the number of successful and unsuccessful trials, to ensure that the models could predict both classes with equal proficiency. To do so, we examined the angular error of each shot relative to the pocket, and removed the *failure* trials closest to the success region, or the worst successful trials - farthest from the centre of the pocket. This selection was based on the expectation that any brain signals related to success feedback would be more prominent for shots greatly differing from the opposite class. To ensure sufficient distinction between trials belonging to different classes, we considered failures that were at least 2 degrees off from the success area. Additionally, we removed the first trial of each block, as the preceding trial was not performed immediately before it. One participant was excluded from the study due to the too low number of trials after the EEG preprocessing pipeline.

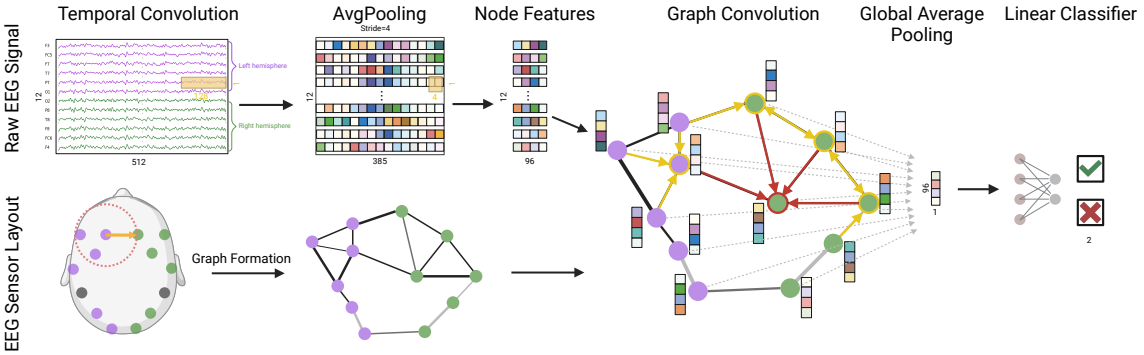

Figure 1: An Illustration of our Graph Neural Network Architecture.

Graph Formation and Model Architecture To model EEG data as brain graphs, we constructed an adjacency matrix $\mathbf{A}$, where $A_{i,j}$ is each term of the matrix that indicates the connection strength between EEG channels i and j. This strength is computed based on spatial proximity, with closer channels having stronger connections. We used a ball tree algorithm to assign channels within a predefined radius as neighbours and calculate the edge weights inversely proportional to the squared distance between channel coordinates, setting the maximum distance between neighbours to 75% of the head radius.

Given the relatively small size of our dataset and the limited number of channels, we designed a shallow temporal CNN and GNN architecture to mitigate the risk of overfitting (see Figure 1). The model processes raw EEG signals with a 1D convolution to capture temporal features, followed by batch normalisation, Parametric ReLU activation, average pooling, and dropout. These features are then input into a Chebyshev graph convolution layer (Defferrard et al., 2016), which propagates information across the graph based on the adjacency matrix $\mathbf{A}$. The output is globally averaged and passed through a linear classification layer. The model is trained with a cross-entropy loss function, using class weights, balancing accurate classification with regularisation. The detailed methodology for constructing the graph representation and the complete neural network architecture are provided in Appendix B.

GNN Spatial Explainability To investigate the differences between the models applied to the four groups, we implemented an explainability pipeline that seeks to identify the most salient channels contributing to the model predictions (Han et al., 2023b). We implemented node and edge masking techniques based on the *GNNExplainer* method (Ying et al., 2019) to investigate the spatial characteristics of our biomarker. This approach involved freezing the weights of the trained GNN model and applying learnable node and edge masks to the raw data and the GNN's edges, respectively. The sigmoid function was utilised to encourage binary masking, enhancing the model's ability to highlight significant spatial features during the biomarker identification process.

The learnable masks were trained using a loss function (1) which combines classification edge mask size loss L_{EMS}, node mask entropy loss L_{NME}, edge mask entropy loss L_{EME}, cross-entropy loss L_{CES}, and node mask size loss L_{NMS}.

$$L_{total} = L_{EMS} + L_{NME} + L_{EME} + L_{CES} + L_{NMS} \tag{1}$$

The classification loss ensured the masked model maintained the original classification accuracy, while the size and entropy losses promoted the masks to be sparse and discrete. The importance score of a node or edge was determined by the corresponding mask value after optimisation, as higher mask values indicated higher importance for the model to minimise the classification loss.

The purpose of this explainability technique is to identify the most salient EEG channels that contribute significantly to the model's predictions. By applying learnable node and edge masks to the raw data and the graph neural network's edges, we can determine the importance of each spatial feature in the classification process. The goal is to find the minimal set of nodes that can be removed while maintaining comparable classification accuracy, thus highlighting the most influential channels for the model's decisions.

Distance Measures between Maps To compare the similarity or divergence between the contribution maps, we utilised a modified version of the Gromov-Wasserstein distance, known as the Sliced Gromov-Wasserstein (SGW) distance, which is less intensive computationally than the Gromov-Wasserstein distance, while preserving the key advantages of the original metric (Vayer et al., 2019). The SGW distance between two graphs is defined as

$$d_{SGW}(G_1, G_2) = \mathbb{E}_{\theta \sim \mathcal{U}(S^{d-1})} \left[W_1 \left(p_\theta(G_1), p_\theta(G_2) \right) \right] \tag{2}$$

where G_1 and G_2 are the two graphs being compared, $p_\theta(G)$ represents the projection of graph G onto the random direction θ, which is sampled uniformly from the unit sphere S^{d-1}. $W_1(p_\theta(G_1), p_\theta(G_2))$ is the 1D-Wasserstein distance between the projected distributions $p_\theta(G_1)$ and $p_\theta(G_2)$. The final SGW distance is the expected value of these 1D-Wasserstein distances over all the random directions θ.

To define the maps for comparison, we transformed the Cartesian electrode coordinates into polar coordinates and added the mask value as the third dimension. This approach allowed us to consider both the mask differences and the spatial information when comparing the maps. To assess the statistical significance of the differences observed between the contribution maps across the various conditions, we performed Mann-Whitney U tests. This non-parametric test was chosen as it does not require the data to follow a normal distribution. In case of paired data, when the samples derived from the same underlying distribution, we performed paired Wilcoxon signed-rank tests.

Training and Evaluation To improve model performance, and based on the assumption of shared underlying features across the groups, we implemented a pre-training strategy prior to the main model training, technique shown to be effective in boosting performances (Hu et al., 2019). Although pretraining is typically defined as training the model on a separate task (Erhan et al., 2010), the uniqueness of the experimental data led us to instead pretrain each group on the other groups. Specifically, we implemented two pretraining approaches: one based on the *Pocket*, and the other on the *Round*. That is, we pretrained each group for 200 epochs on the two other groups belonging to the opposite pocket side or the opposite round, in order to have two comparable groups in every pretraining scenario. This pretraining approach allowed us to leverage the shared features across the groups, potentially enhancing the model's ability to generalise and improve its performance.

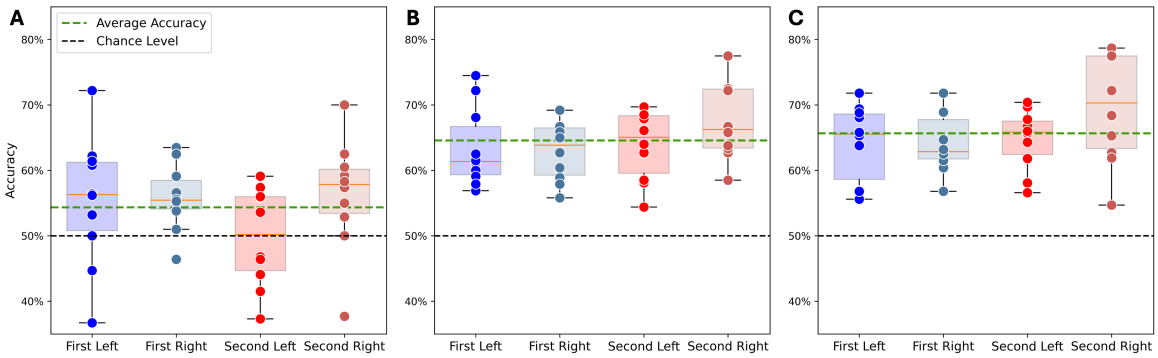

Figure 2: Performances of GNN models with (**A**) no pretraining, (**B**) pretraining on opposite *Round*, (**C**) and pretraining on opposite *Pocket*. The 10-fold cross validation accuracy distributions of the four groups are displayed separately for each pretraining mode. Black dashed line represents chance level, while the green dashed line shows the average accuracy across groups for the individual mode.

Following pretraining, a 10-fold cross-validation training procedure was then performed. The split in folds was done ensuring that each participant was evenly represented across the folds, while also maintaining a balance between the classes, due to the individual variability in brain data. In this case, the model was trained for only 150 epochs due to the smaller sample size. Both models were trained using a `batch size` of 32 and a `learning rate` of 1e-4 to ensure a trade-off between optimality and training costs. To prevent overfitting, we implemented L2-optimisation in the `Adam` optimiser, edge dropout in the initial graph ($p = 0.2$), and dropout at the end of the temporal convolution layer ($p = 0.35$) in all the models (Zhang, 2018; Rong et al., 2019; Srivastava et al., 2014). To evaluate the models performances, we examined accuracy, precision, recall, and F1-score on the validation set, which was not used for any training decisions or hyperparameter optimisation, to assess the models' ability to generalise and perform on unseen data.

3. Results

Model Performance The pretrained models, based on the *Round* or *Pocket* groupings, led to significant improvements in model performance across all evaluation metrics. Without pretraining, our GNN models struggled to extract meaningful patterns from the EEG data, achieving an average accuracy of only 0.54 across all groups, barely above chance level. The introduction of pretraining, however, led to a significant boost in model performance. Both *Round*-based (Figure 2B) and *Pocket*-based (Figure 2C) pretraining resulted in substantial improvements, with average accuracies across groups increasing to 65% and 66% respectively, with values up to 70% in the individual groups. This consistent enhancement across all participant groups suggests that pretraining enables the GNNs to learn more robust and generalisable features from the EEG data. The full metrics comparison is shown in Appendix C.

Explainability Results The explainability analysis of our GNN models revealed consistent group-specific neural signatures across different pretraining conditions. While the

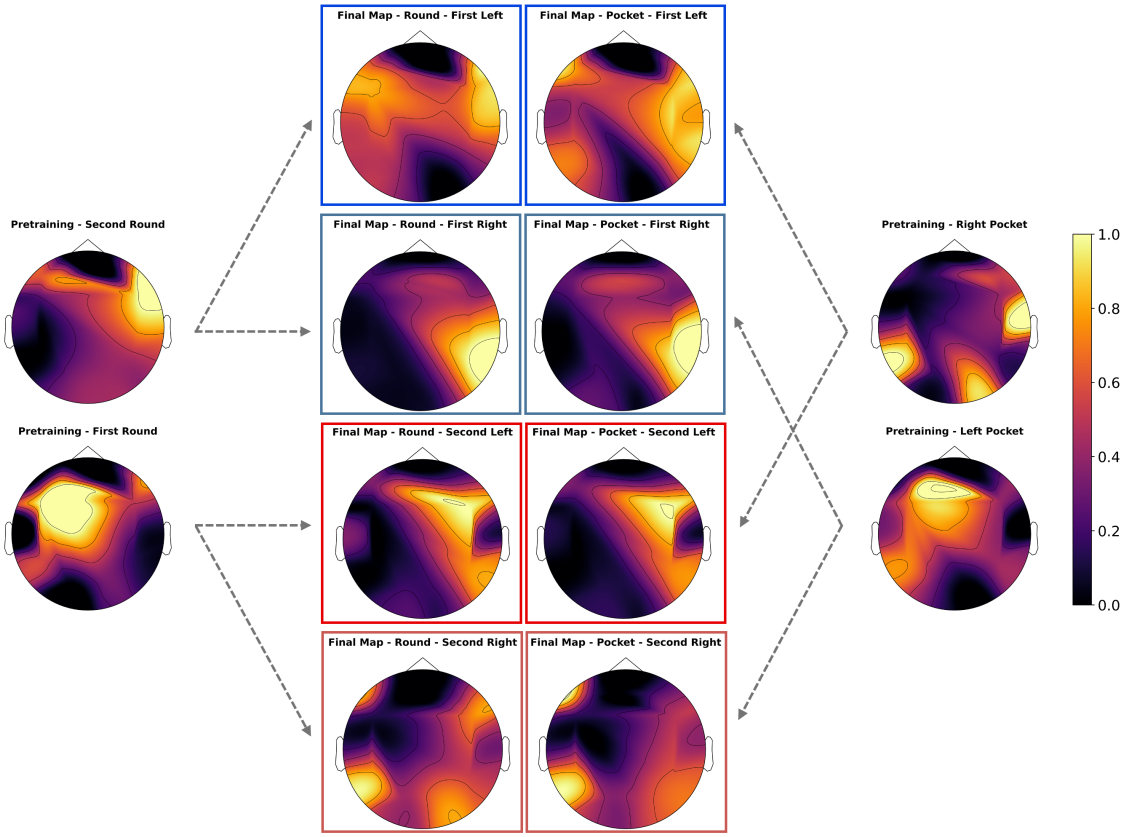

Figure 3: Brain activation maps across the four groups (*First Left, First Right, Second Left, Second Right*). Arrows connect the pretraining maps used, based on opposite round (first column) or opposite pocket (fourth column) to the final maps of the four groups.

Round or *Pocket* pretraining maps are very different, the brain maps for each group exhibited similar spatial patterns of importance, irrespective of the pretraining method (Figure 3). This consistency suggests that the models have identified stable neural representations for each group. The persistence of these group-specific patterns across pretraining conditions indicates the presence of underlying neural architectures unique to each group's learning experience. This finding implies potential differences in the effect of reinforcement learning from the previous trial based on aspects of the task. To quantify the similarities between these neural representations, we measured the Sliced Gromov-Wasserstein (SGW) distances (Figure 4A), which showed the highest similarities between the four groups to themselves across the different pretraining conditions. This structure provides evidence for the stability of group-specific neural patterns, suggesting that the characteristics of each group's brain activity during the task remain relatively consistent and robust to varying training conditions of the models.

Comparing the distances of each group's pretraining map to their final map (Figure 4B - II-III) and the distances between the final maps of the groups (Figure 4B - IV), it is evident that the GNNs primarily relied on the group-specific spatial information rather

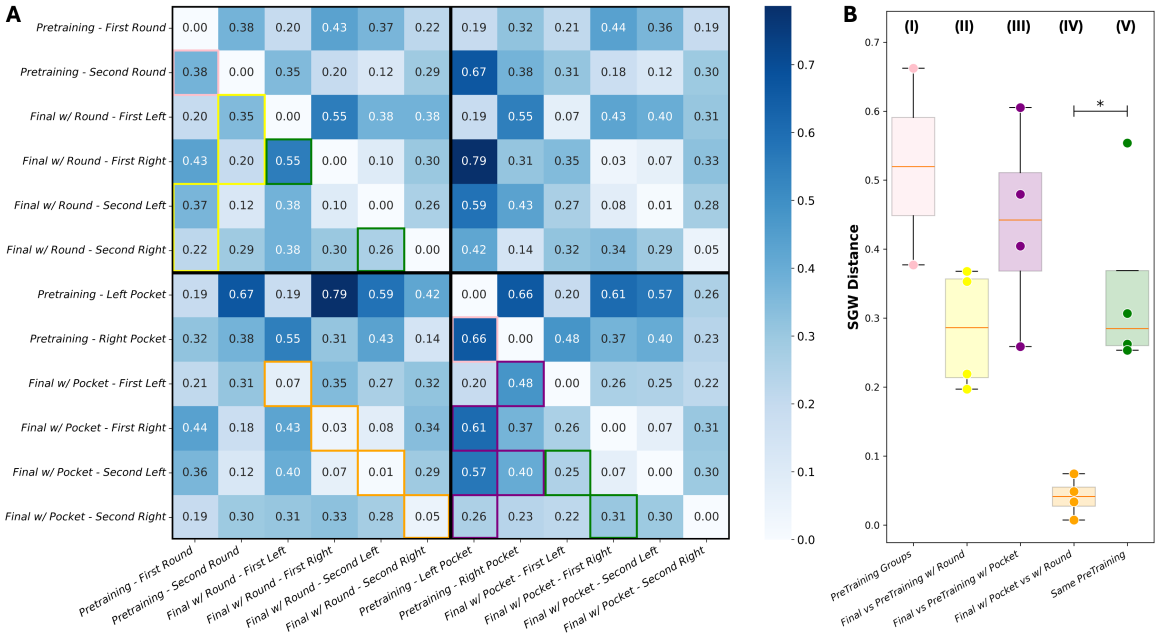

Figure 4: **(A)** Representation Distance Matrix between group maps **(B)** Distribution of distances between groups of: (I) same pretraining strategies; final vs initial maps when pretraining with opposite (II) round and (III) pocket; (IV) final maps within group with different pretraining; final maps between groups that had the same pretraining maps as starting point (V). The quantities reported in the boxplots are framed in the RDM using the same colour scheme.

than the underlying common information. Similarly, the significant difference ($p = 0.03$) between the distances of groups starting from the same pretraining (Figure 4B - V) and the distances between the final maps (Figure 4B - IV) represents the distinctiveness of each group's neural representations, suggesting the minor importance of the pretraining strategy. Noticeable distributional differences were found between the remaining features, although these differences were not statistically significant due to the small sample size.

These results suggest that despite the complex nature of brain activity during motor tasks, there exist discernible, group-specific patterns that can be identified by machine learning models. The ability of the models to capture these consistent patterns across different pretraining strategies demonstrates the adaptability of the GNN architecture in processing complex EEG data, suggesting that these models can extract relevant features from the spatial relationships between EEG channels, providing insights into the neural basis of motor learning and feedback processing.

4. Discussion

This study used GNNs to understand the neural correlates of performance feedback in motor planning in a real-world billiard task. By analysing the spatial relationships in the EEG data, our models revealed consistent group-specific neural signatures, providing insights into the complex brain activity underlying motor skill acquisition. GNNs have shown great

promise in analysing static, well-structured brain data, such as that from motor imagery tasks (Hou et al., 2022; Sun et al., 2022). Previous research has highlighted the ability of GNNs to effectively capture spatial relationships in brain data (Kim et al., 2021; Azevedo et al., 2022). Yet, the successful application of these models to dynamic, real-world tasks remains an active area of research, where novel techniques are needed to overcome the inherent complexities of brain activity in natural environments.

Here we manipulated the GNNs training pipeline, implementing different pretraining strategies using both opposite *Round* or *Pocket*, which led to a significant improvement in model performance, despite the unique and complex nature of the task, indicating that the pretraining process enabled the GNNs to acquire more robust and generalisable features from the EEG data. This consistent enhancement suggests that the pretraining methods allowed the models to better navigate the intricate manifold of neural representations, coherently with previous studies (Hu et al., 2019, 2020). Notably, the comparable performance gains for both pretraining approaches imply that the GNNs were able to effectively leverage different aspects of the experimental design to enhance their predictive capabilities, hinting at the presence of underlying neural patterns that persist across experimental conditions.

Interestingly, persistent group-specific patterns were observed in the brain maps generated through our explainability analysis. The GNN models exhibited robust spatial invariance, and consistently identified distinct spatial activation patterns for each of the four groups, regardless of the pretraining approach. This consistency of group-specific neural representations suggests that individual differences in motor learning strategies or feedback processing mechanisms are reflected in stable, identifiable patterns of brain activity.

Taken together, these results suggest that the neural correlates underlying motor learning and performance feedback processing are influenced by both individual differences in learning strategies as well as the specific demands and characteristics of the task. The consistent detection of group-specific neural representations by the GNN models, irrespective of the pretraining method, underscores the robustness of our approach in extracting meaningful neural signatures associated with complex cognitive and motor processes.

Limitations and Future Work This study provides valuable insights into the neural correlates of motor learning using GNNs, but several limitations exist. The small sample size restricts the generalisability of the findings and may have contributed to sub-optimal model performance, suggesting the need to increase the sample size in future studies which could enhance accuracy and robustness. Individual differences in brain activity and learning strategies also remain a challenge, suggesting the need for more personalised modelling approaches and the integration of additional transfer learning techniques. Further work is needed to connect the spatial patterns in EEG to specific cognitive processes and neurophysiological mechanisms, as well as investigate the temporal aspects of motor planning. Lastly, the study focused on short-term feedback, leaving open the question of how neural representations change over longer periods of skill acquisition.

Conclusion This study highlights the potential of Graph Neural Networks to provide meaningful insights into complex real-world neural data. The consistent group-specific spatial patterns found underscore the robustness of this method in capturing differences in the underlying cognitive and motor processes. This study lays the groundwork for future advancements in understanding complex neural representations using GNNs.

References

Scott T Albert and Reza Shadmehr. The neural feedback response to error as a teaching signal for the motor learning system. *Journal of Neuroscience*, 36(17):4832–4845, 2016.

Tiago Azevedo, Alexander Campbell, Rafael Romero-Garcia, Luca Passamonti, Richard AI Bethlehem, Pietro Lio, and Nicola Toschi. A deep graph neural network architecture for modelling spatio-temporal dynamics in resting-state functional mri data. *Medical Image Analysis*, 79:102471, 2022.

S. Berti. Cognitive control after distraction: event-related brain potentials (erps) dissociate between different processes of attentional allocation. *Psychophysiology*, 45 4:608–20, 2008. doi: 10.1111/j.1469-8986.2008.00660.x.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

James C Eliassen, Martine Lamy, Jane B Allendorfer, Erin Boespflug, Daniel P Bullard, Matthew S Smith, Jing-Huei Lee, and Stephen M Strakowski. Selective role for striatal and prefrontal regions in processing first trial feedback during single-trial associative learning. *Brain research*, 1458:56–66, 2012.

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

Shlomi Haar and A Aldo Faisal. Brain activity reveals multiple motor-learning mechanisms in a real-world task. *Frontiers in human neuroscience*, 14:354, 2020.

Shlomi Haar, Guhan Sundar, and A Aldo Faisal. Embodied virtual reality for the study of real-world motor learning. *Plos one*, 16(1):e0245717, 2021. doi: 10.1371/journal.pone.0245717.

Jinpei Han, Xiao Gu, Guang-Zhong Yang, and Benny Lo. Noise-factorized disentangled representation learning for generalizable motor imagery eeg classification. *IEEE Journal of Biomedical and Health Informatics*, 2023a.

Jinpei Han, Xiaoxi Wei, and A Aldo Faisal. Eeg decoding for datasets with heterogenous electrode configurations using transfer learning graph neural networks. *Journal of Neural Engineering*, 20(6):066027, 2023b.

Diana Henz and Wolfgang I Schöllhorn. Differential training facilitates early consolidation in motor learning. *Frontiers in behavioral neuroscience*, 10:199, 2016.

Yimin Hou, Shuyue Jia, Xiangmin Lun, Ziqian Hao, Yan Shi, Yang Li, Rui Zeng, and Jinglei Lv. Gcns-net: a graph convolutional neural network approach for decoding time-resolved eeg motor imagery signals. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1857–1867, 2020.

Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, 34:4314–4327, 2021.

C Kothe. The artifact subspace reconstruction method. *Accessed: Jul*, 17:2017, 2013.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Caroline Di Bernardi Luft. Learning from feedback: The neural mechanisms of feedback processing facilitating better performance. *Behavioural brain research*, 261:356–368, 2014.

Y Narasimhulu, Ashok Suthar, Raghunadh Pasunuri, and China Venkaiah Vadlamudi. Ckd-tree: An improved kd-tree construction algorithm. In *ISIC*, pages 211–218, 2021.

Federico Nardi, Aldo A Faisal, and Shlomi Haar. Motor learning mechanisms are not modified by feedback manipulations in a real-world task. *bioRxiv*, 2024.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38:5391–5420, 2017.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Biao Sun, Han Zhang, Zexu Wu, Yunyan Zhang, and Ting Li. Adaptive spatiotemporal graph convolutional networks for motor imagery classification. *IEEE Signal Processing Letters*, 28:219–223, 2021.

Biao Sun, Zhengkun Liu, Zexu Wu, Chaoxu Mu, and Ting Li. Graph convolution neural network based end-to-end channel selection and classification for motor imagery brain–computer interfaces. *IEEE transactions on industrial informatics*, 19(9):9314–9324, 2022.

D. Talsma and M. Woldorff. Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience*, 17: 1098–1114, 2005. doi: 10.1162/0898929054475172.

Huiling Tan, Ned Jenkinson, and Peter Brown. Dynamic neural correlates of motor error monitoring and adaptation during trial-to-trial learning. *Journal of Neuroscience*, 34(16): 5678–5688, 2014.

Jonathan Tsay, Nathan Steadman, Melanie Fleming, Mareike Gann, Irene Di Giulio, Cosima Graef, Jinpei Han, Kavindu Jayasinghe, Matthew Mitchell, Robert Chen, et al. Bridging the gap between experimental control and ecological validity in human sensorimotor science. *The Journal of physiology*, 602(17):4085–4087, 2024.

Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty. Sliced gromov-wasserstein. *arXiv preprint arXiv:1905.10124*, 2019.

Ping Wang, Aimin Jiang, Xiaofeng Liu, Jing Shang, and Li Zhang. Lstm-based eeg classification in motor imagery tasks. *IEEE transactions on neural systems and rehabilitation engineering*, 26, 2018.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Yuan Yuan, Hongwei Mao, and Jennie Si. Cortical neural responses to previous trial outcome during learning of a directional choice task. *Journal of neurophysiology*, 113(7): 1963–1976, 2015.

Manuel Zaepffel, Romain Trachel, Bjørg Elisabeth Kilavik, and Thomas Brochier. Modulations of eeg beta power during planning and execution of grasping movements. *PloS one*, 8(3):e60060, 2013.

Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018.

Appendix A. Experimental Setup and Design

A.1. EVR Setup

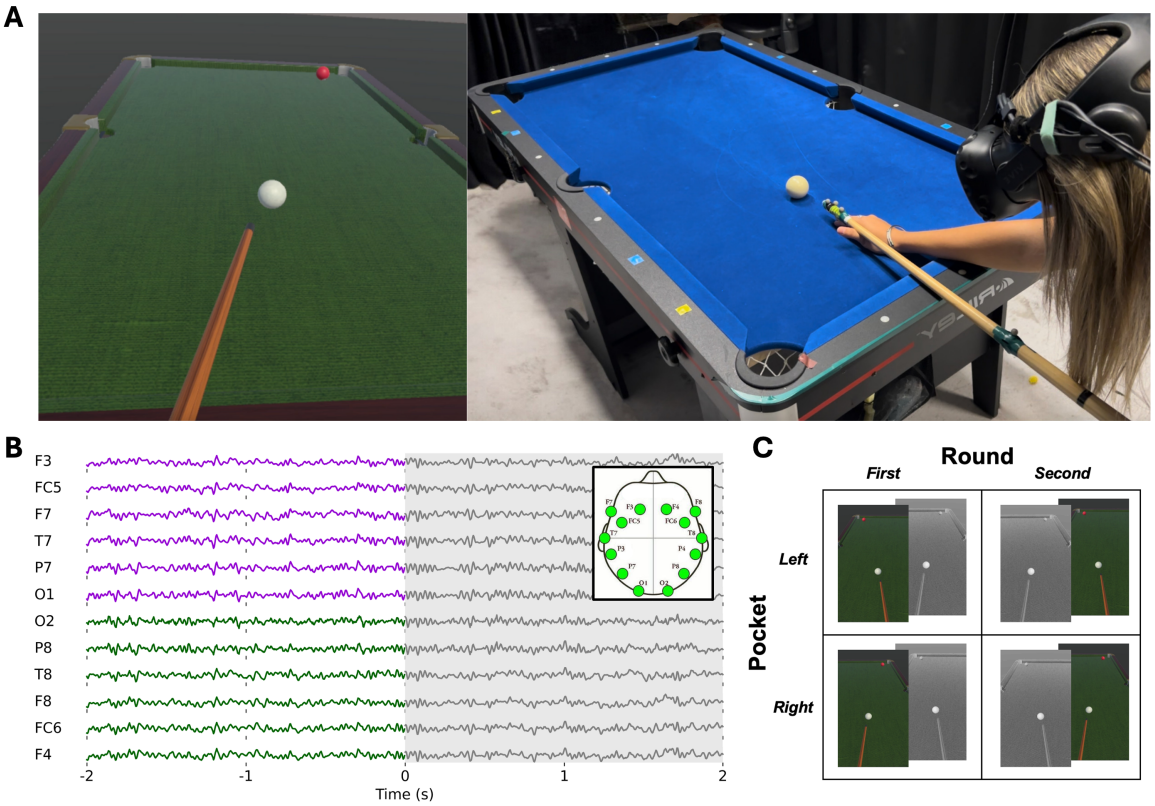

Figure 5: (**A**) Billiard task in the Embodied Virtual Reality setup: participants playing pool in real-world (*right*) are immersed in the VR environment in which visual feedback manipulations are applied (*left*); (**B**) 12-channels EEG data sample used for the GNN models: out of a window of 4 seconds around each trial, only the 2s pre-shot window is considered for the analysis; (**C**) Experimental conditions of the four groups: combinations of *Left* and *Right* pockets and *First* and *Second* rounds were considered according to the session and shooting direction experienced by the participants.

A.2. Label Derivation

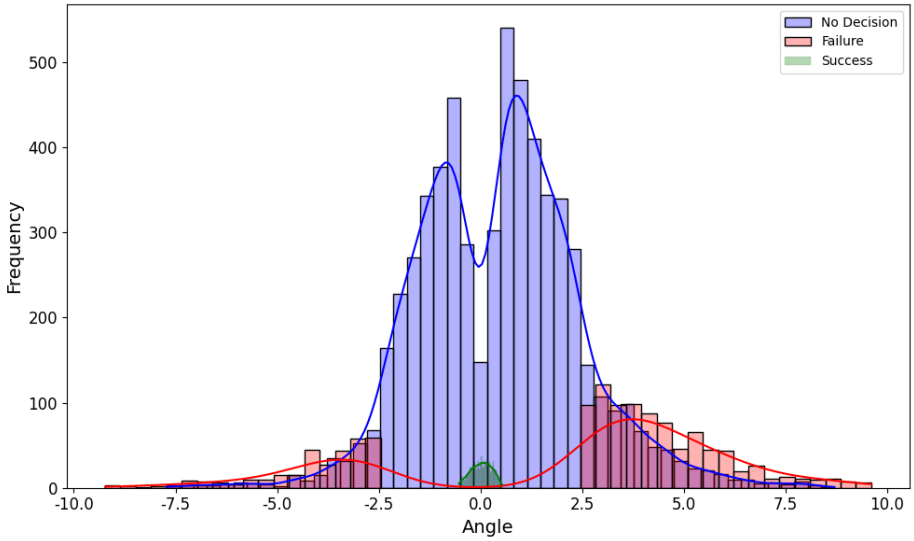

Figure 6: Distribution of all trials of all participants over the standardised angle range (0° represents a perfect successful shot), divided by label. The *No Decision* category includes the failure trials which are too close to the success funnel (in green), according to a pre-determined threshold of 2 degrees. $N = 39$.

Appendix B. Graph Formation and Model Architecture

B.1. Graph Formation

To represent our EEG data as brain graphs, we first construct an adjacency matrix $\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{C \times C}$, with C denoting the total number of EEG channels (nodes). The elements of the matrix, $A_{i,j}$, represent the connection strength or edge weight between nodes i and j, which in our case correspond to the spatial proximity between EEG channels. A stronger connection between two nodes results in a higher edge weight, while an absence of connection is indicated by $A_{i,j} = 0$.

Given the spatial coordinates of the EEG channels, we employ a ball tree algorithm, called the `cKDTree`, to efficiently identify neighbouring channels within a predefined radius r (Narasimhulu et al., 2021). For each channel i, the algorithm identifies all channels j within this radius, forming potential edges between them. We empirically found that $r = 75\%$ head radius was the best trade-off between a low number of neighbours for each node, to avoid misleading information coming from very far channels, and the inclusion of all the channels in the headset.

The weight of the edge between nodes i and j is then computed as the inverse square of the Euclidean distance between their corresponding coordinates. Mathematically, the weight $w_{i,j}$ of the edge is given by:

$$w_{i,j} = \begin{cases} \frac{1}{\|\mathbf{x}_i - \mathbf{x}_j\|^2} & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\| \neq 0, \\ 0 & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\| = 0, \end{cases} \tag{3}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the spatial coordinates of channels i and j, respectively, and $\|\mathbf{x}_i - \mathbf{x}_j\|$ denotes the Euclidean distance between these channels. The inverse square relationship emphasizes stronger connections between closer nodes while diminishing the influence of more distant nodes. Self-links (i.e., $A_{i,i}$) are avoided to ensure the adjacency matrix strictly represents inter-channel relationships.

The resulting adjacency matrix $\mathbf{A}$ thus captures the spatial structure of the EEG channels, supported by neurophysiological findings that suggest connective strength diminishes with squared distance.

B.2. GNN Architecture

We designed the model to have a shallow temporal CNN and GNN architecture, in order to mitigate the risk of overfitting due to the relatively small sample size. Our model takes in two primary inputs: the raw EEG signals, represented as node features with dimensions $C \times T$, where C denotes the number of channels (nodes) and T signifies the number of time points, and the structural information of the graph, represented by the $C \times C$ adjacency matrix $\mathbf{A}$, which encodes the connectivity and edge weights among the nodes.

The architecture begins by applying a one-dimensional convolution (1D-CNN) along the temporal axis of the raw EEG signals. The convolution operation uses a kernel size of $\frac{f_s}{2}$, where f_s is the sampling rate of the EEG signals (Lawhern et al., 2018). This kernel size is chosen to capture the relevant frequency information in the EEG signals effectively. The temporal convolution layer is defined as:

$$\mathbf{H}_1 = \text{PReLU}(\mathbf{W}_{\text{conv}} * \mathbf{X} + \mathbf{b}_{\text{conv}}), \tag{4}$$

where $\mathbf{X} \in \mathbb{R}^{C \times T}$ is the input EEG data, $\mathbf{W}_{\text{conv}}$ is the convolutional kernel, $\mathbf{b}_{\text{conv}}$ is the bias term, and $*$ denotes the convolution operation along the time axis. After the convolution, we apply batch normalisation and a Parametric ReLU (PReLU) activation function to introduce non-linearity, followed by average pooling and dropout layers. The average pooling layer reduces the temporal dimensionality, thereby lowering the risk of overfitting in subsequent layers.

Following temporal feature extraction, the reduced node features are fed into a single Chebyshev graph convolution layer to propagate information across neighbouring nodes (EEG channels). The Chebyshev graph convolution is defined as:

$$\mathbf{H}_2 = \sum_{k=0}^{K-1} \mathbf{T}_k(\tilde{\mathbf{L}})\mathbf{H}_1\mathbf{\Theta}_k, \tag{5}$$

where $\tilde{\mathbf{L}} = 2\mathbf{L}/\lambda_{\max} - \mathbf{I}_C$ is the rescaled Laplacian matrix derived from the adjacency matrix $\mathbf{A}$, $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the graph Laplacian, $\mathbf{D}$ is the degree matrix, $\mathbf{T}_k(\cdot)$ represents the Chebyshev polynomial of order k, $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{L}$, and $\mathbf{\Theta}_k$ are the learnable parameters of the layer. This formulation allows for localised spectral filtering, enabling the model to capture complex spatial dependencies among EEG channels while maintaining computational efficiency.

The output from the Chebyshev GNN layer, $\mathbf{H}_2$, undergoes a *Softplus* activation to ensure non-negative activations. Subsequently, global average pooling is applied across all node features to aggregate the information into a single feature vector. Finally, the aggregated feature vector $\mathbf{h}_{\text{global}}$ is passed through a linear classification layer to perform the final prediction. Overall, all layers in the architecture are trained to minimise a class-weighted cross-entropy loss function, defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log(\hat{y}_{n,c}), \tag{6}$$

where N is the number of samples, C is the number of classes, $y_{n,c}$ is the true label of class c for sample n, and $\hat{y}_{n,c}$ is the predicted probability for class c after applying a softmax function to the output of the linear classification layer.

This simple architecture efficiently leverages both temporal and spatial information from the EEG data, allowing for accurate classification while minimising the risk of overfitting.

Appendix C. Model Performances

	No Pretraining				**Pretraining w/ Round**				**Pretraining w/ Pocket**			
	Acc	*Prec*	*Rec*	*F1*	*Acc*	*Prec*	*Rec*	*F1*	*Acc*	*Prec*	*Rec*	*F1*
First Left	0.55	0.55	0.57	0.55	0.63	0.61	0.74	0.67	0.64	0.61	0.80	0.69
First Right	0.56	0.53	0.57	0.54	0.63	0.63	0.65	0.64	0.64	0.64	0.66	0.65
Second Left	0.50	0.50	0.42	0.44	0.64	0.67	0.55	0.60	0.65	0.64	0.67	0.66
Second Right	0.56	0.58	0.53	0.54	0.68	0.69	0.65	0.67	0.70	0.71	0.68	0.69
Average	0.54	0.54	0.52	0.52	0.65	0.65	0.65	0.65	0.66	0.65	0.70	0.67

Table 1: GNN Model Performances for the four groups, obtained as the average of the performances on the validation sets of the 10-fold cross validation. The three groups of columns show metrics without any pretraining, pretraining on opposite *Round* and opposite *Pocket*, respectively. The last row shows the average for the metrics across the four groups.
Acc: Accuracy; *Prec*: Precision; *Rec*: Recall; *F1*: F1-score.

Appendix D. GNN Explainability Maps - Evolution over Learning

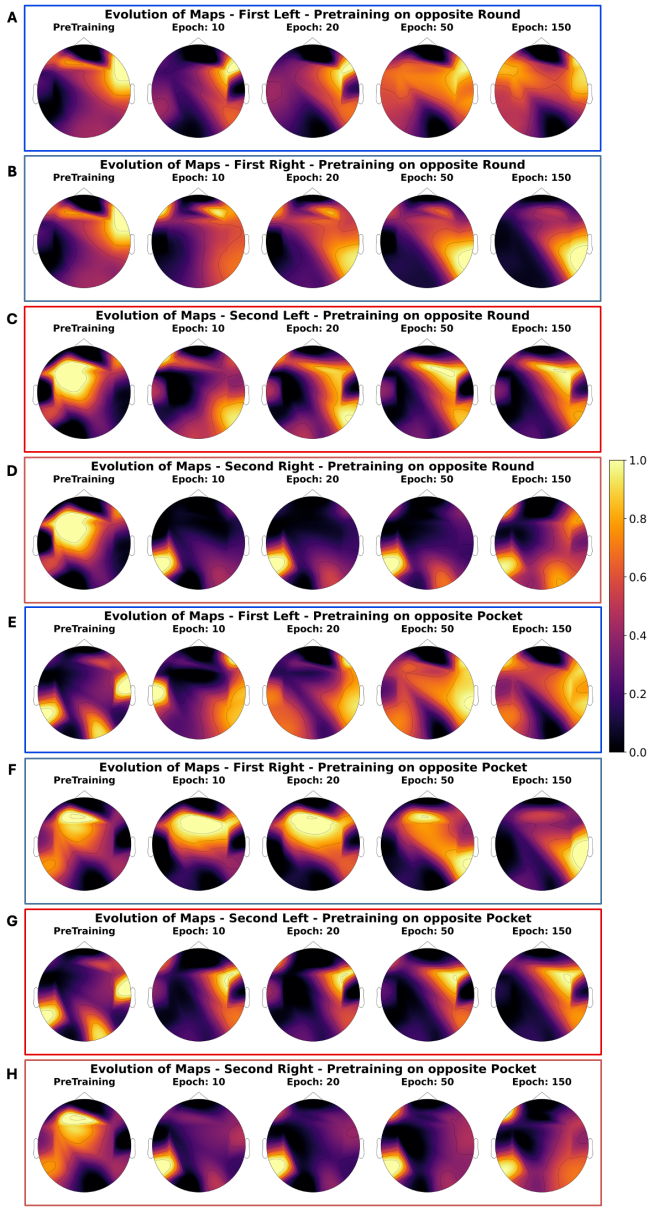

Figure 7: Brain Maps Evolution across learning when pretraining on opposite *Round* (A-D) and *Pocket* (E-H), for the *First Left* (A,E), *First Right* (B,F), *Second Left* (C,G), and *Second Right* (D, H) groups. The five columns show the explainability maps for pretraining, epochs 10, 20, 50 and 150 respectively. The colour bar refers to all the plots and represents the standardised contribution of the different channels, according to explainability methods described in Section 2.