# DyNeMoC: A SEMI-SUPERVISED ARCHITECTURE FOR CLASSIFYING TIME SERIES BRAIN DATA

**Abu Mohammad Shabbir Khan**[1 2 *] **Chetan Gohil**[3] **Pascal Notin**[2]
**Joost van Amersfoort**[2] **Mark Woolrich**[3] **Yarin Gal**[2]
[1]Relation Therapeutics
[2]Oxford Applied and Theoretical Machine Learning (OATML), University of Oxford
[3]Oxford Centre for Human Brain Activity (OHBA), University of Oxford

## ABSTRACT

Understanding how different regional networks of the brain get activated and how those activations change over time can help in identifying the onset of various neurodegenerative diseases, studying the efficacy of different treatment regimens for those illnesses, and developing brain-computer interfaces for patients with different types of disabilities. To explain dynamic brain networks, an RNN-VAE model named DyNeMo has recently been proposed. This model can take into account the whole recorded history of brain states while modeling their dynamics and is able to better capture the complexities in larger datasets than previous works. In this paper, we show that the latent representations learned by DyNeMo through unsupervised training are not sufficient for downstream classification tasks and propose a new semi-supervised model named DyNeMoC that overcomes this shortcoming. The downstream task we study is the classification of visual stimuli from MEG recordings. We show that both of our proposed variants of DyNeMoC — DyNeMoC-RNN and DyNeMoC-Transformer — lead to more useful latent representations for stimuli classification with the transformer variant outperforming the RNN one. Learning representations that are directly linked to a downstream task in this manner could ultimately be used to improve the monitoring and treatment of certain neurodegenerative diseases and building better brain-computer interfaces.

## 1 INTRODUCTION

Brain states are regional networks of neurons which spontaneously activate while at rest (Biswal et al., 1995; Fox & Raichle, 2007; Raichle et al., 2001) and while performing various cognitive tasks (Kurth-Nelson et al., 2015; Isik et al., 2014; Carlson et al., 2011). A useful imaging modality for studying the dynamic nature of these brain states is MEG (Lopes da Silva, 2013) because it provides a direct measure of neuronal activity at a millisecond resolution — a highly desirable property for studying brain activities at their natural time scale (Proudfoot et al., 2014).

Characterizing the spatio-temporal dynamics of brain states can not only help us in gaining a better understanding of the underpinnings of cognition (Buzsáki & Draguhn, 2004; Bressler & Menon, 2010) but also has numerous healthcare applications. Functional connectivity (Friston, 1994) of brain states are being used to study the diagnosis (Josef Golubic et al., 2017; Schoonhoven et al., 2019; Dimitriadis et al., 2018; Schumacher et al., 2019; Fiorenzato et al., 2019; Dopper et al., 2014; Mandal et al., 2018; Babiloni et al., 2020) of and intervention (Shigihara et al., 2020b;a) for different neurodegenerative diseases. Correct identification of brain states can also lead to better brain-computer interfaces with implications in patient-care (Liberati et al., 2012; Mudgal et al., 2020).

Historically, sliding window techniques have been used to infer dynamic brain networks from neuroimaging data (Wendling et al., 2009; Allen et al., 2012). More recently, there has been a shift to using unsupervised learning approaches, such as Hidden Markov models (HMMs) (Baker et al., 2014; Vidaurre et al., 2016; 2017; 2018), and variational autoencoders (VAEs) (Perl et al., 2020)

*Work done while at OATML. Correspondence to shabbir@relationrx.com

for this type of work. One recent development along this line has been the introduction of an unsupervised RNN-VAE (Chung et al., 2015; Bowman et al., 2015; Fabius & Van Amersfoort, 2014) named DyNeMo (Gohil et al., 2022) which is a generative model and Bayesian inference scheme for identifying brain networks.

In this work, we devise a semi-supervised architecture named DyNeMoC that builds on DyNeMo. We propose two variants of DyNeMoC — DyNeMoC-RNN and DyNeMoC-Transformer — and conduct experiments on a real-world MEG dataset that contains neural responses to visual stimuli. We establish that DyNeMoC is better than DyNeMo for learning latent state representations that are useful for certain downstream tasks of interest such as the classification of visual stimuli. We further demonstrate that the DyNeMoC-Transformer architecture can far outperform DyNeMoC-RNN in this aspect. We hypothesize the representations learned by DeNeMoC would also be useful in other applications if the same brain states are recruited.

## 2 BACKGROUND

DyNeMo is a VAE (Kingma & Welling, 2013; 2019) with a bidirectional RNN encoder, a unidirectional RNN prior network, and a decoder which is further comprised of state means and covariances. The core modeling assumption behind it is that neural time series data is generated from a finite set of $J$ latent brain states, where each state can be represented by a distinct multivariate normal distribution. These states probabilistically mix, i.e. linearly combine with each other with coefficients $\alpha_{jt}$ at each time step $t$ (such that $\alpha_{jt} \in [0, 1]$ and $\sum_{j=0}^{J} \alpha_{jt} = 1$) to generate a time-varying description of the means and covariances of the data. Here, $\alpha_{jt}$ for each state at time step $t$ is computed by using the softmax operation on the posterior logits (latent representation) $\boldsymbol{\theta}$ produced by the encoder.

Gohil et al. (2022) trained DyNeMo by minimizing the variational free energy, $\mathcal{L} = -LL + KL$ , where $-LL$ denotes the negative log-likelihood (NLL) of an observation being generated from the learned state means and covariances, and $KL$ denotes KL divergence from the prior to the posterior distribution. Further details regarding DyNeMo can be found in Gohil et al. (2022).

## 3 METHODOLOGY

In this work, we evaluate the usefulness of the latent representations learned by DyNeMo in a downstream classification task. We propose these latent representations can be improved by jointly training a multilayer perceptron (MLP) classifier with DyNeMo. This approach incentivizes the model to encode information useful for the downstream task (class labels) into the latent representation. We call this model DyNeMoC and provide its general architecture and data flow in Figure 1.

The MLP classifier of DyNeMoC is fed the inferred logit courses $\boldsymbol{\theta}$ by the encoder as a flattened vector. We used inferred $\boldsymbol{\theta}$ courses instead of inferred $\boldsymbol{\alpha}$ courses here because unlike $\alpha_{jt}$ which are confined between 0 and 1, $\theta_{jt}$ could take any real value.

Our updated loss function $\mathcal{L}_u$ for the joint network thus became as follows:

$$\mathcal{L}_u = -LL + KL + w_c \times CC \tag{1}$$

where $CC$ is the cross entropy between the actual and predicted labels, and $w_c$ is a hyperparameter controlling the weight of the cross-entropy loss.

Gohil et al. (2022) originally devised DyNeMo with RNNs, particularly LSTMs (Hochreiter & Schmidhuber, 1997). We label a DyNeMoC model stemming from the original architecture as a DyNeMoC-RNN. Recognizing the RNNs are essentially sequence-learning networks, we further created an enhanced version of DyNeMoC based on transformers (Vaswani et al., 2017) which we call DyNeMoC-Transformer. In this architecture, we used a small-scale BERT (Devlin et al., 2018) model for the encoder and a small-scale GPT-2 (Radford et al., 2019) model for the prior network. We opted for small-scale BERT and GPT-2 models because of limited amounts of labelled training data being available.

The DyNeMo component of the first DyNeMoC-RNN model we designed was similar to the DyNeMo described in Gohil et al. (2022). This DyNeMoC-RNN model had 2.1M learnable parameters, and we call it DyNeMoC-RNN-Small. Our DyNeMoC-Transformer model, on the other

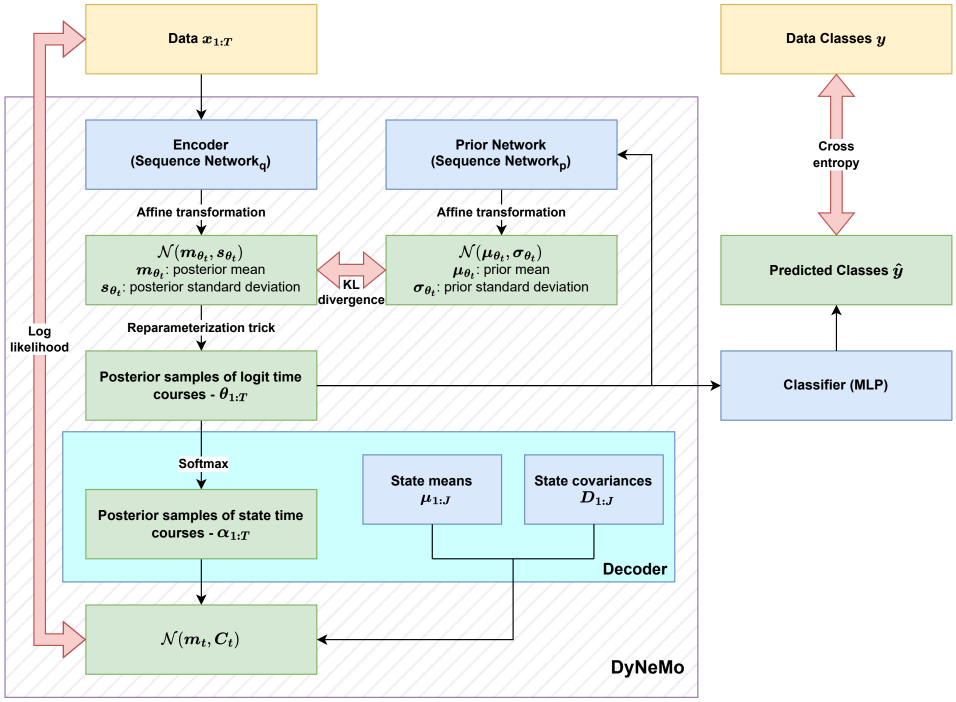


Figure 1: The general architecture and data flow of DyNeMoC

hand, had 9.2M parameters. To do a proper comparison between, DyNeMoC-RNN and DyNeMoC-Transformer, we designed DyNeMoC-RNN-Large with 9.8M parameters. The particular hyperparameters of the different DyNeMoC models are provided in Appendix A.2.

# 4 EXPERIMENTS

## 4.1 DATASET

In this study, we used a dataset created by Cichy et al. (2016) that contained MEG data collected from 15 subjects while they were shown images of 118 different object classes (also referred to as conditions) at 0.9–1s intervals (around 30 times for each object class). We first pre-processed the MEG data by performing band-pass filtering on it and then downsampling it to 250Hz frequency. We kept each trial in the range -0.2–0.6s (inclusive). So each trial in this dataset had $(0.6 - (-0.2)) \times 250 + 1 = 201$ timesteps. We then performed a 70-15-15 train-validation-test split of the dataset while maintaining the relative balance among the trials in all the conditions for all subjects.

The 306 sensors that were used to record the MEG data in this dataset measure different properties of magnetic fields at different scales. Moreover, working with covariance matrices of dimensions $(306 \times 306)$ is also computationally quite expensive. Hence, for each subject, we standardized, PCA-transformed, and re-standardized their trials to reduce the data dimensions and bring features to the same scale. We settled for 80 principal components while conducting PCA as these components explained over 98% of the observed variance in each subject's data on average. We note that it was necessary to process each subject's data separately here because the data among subjects do not align in sensor-space and PCA-space owing to the structural and functional variety in human brains as well as the non-uniformity in sensor placement (Zhang et al., 2017). As such, in each of our experiments, a model was trained and evaluated on one subject only.

Cichy et al. (2016) reported that for the different conditions in the dataset, they observed the earliest onset of significant neuronal activity at 77ms and the last peak point at 326ms. Hence, in all of our experiments, the input to the classifier was the 50–350ms (inclusive) windows of the trials directly or in the form of $\boldsymbol{\theta}$ or $\boldsymbol{\alpha}$ courses.

Table 1: Summarized classification accuracies on the test sets of different subjects

| Subject | SVM Baseline | MLP Baseline | DyNeMo RNN-Small + MLP | DyNeMo RNN-Large + MLP | DyNeMo Transformer + MLP | DyNeMoC RNN-Small | DyNeMoC RNN-Large | DyNeMoC Transformer |
|---|---|---|---|---|---|---|---|---|
| **Mean (Std. Err.)** | 0.20 (0.04) | 0.21 (0.05) | 0.07 (0.02) | 0.06 (0.02) | 0.06 (0.01) | 0.29 (0.07) | 0.30 (0.07) | **0.40 (0.09)** |

### 4.2 Baselines

**SVM and MLP baselines:** We trained a linear SVM (Noble, 2006) (which Cichy et al. (2016) also used) for each subject with a multi-class classification objective (different from the binary classification investigated in Cichy et al. (2016)) by directly feeding it the time window 50–350ms of the trials as input. Furthermore, we trained MLPs with 1024 hidden units, 2 layers, GELU activations (Hendrycks & Gimpel, 2016), and 0.5 dropout (Srivastava et al., 2014) on the same inputs.

**Two-step baselines:** For each hyperparameter configuration of each variant of DyNeMoC, we first trained its DyNeMo component in an unsupervised way and then used the flattened $\alpha_{63:137}$ vectors (corresponding to 50–350ms) to train the MLP components separately. This baseline model was meant to help us understand how useful the latent representations learned by DyNeMo are when it is trained with its original objective.

### 4.3 Results

We present the summary of test accuracies of the different models in Table 1 and the detailed subject-wise test accuracies in Table 4 of Appendix A.4. We first note that the two-step baseline models had very poor accuracies. This established that a completely unsupervised training of DyNeMo cannot lead to latent representations that are effective for downstream classification. This might also indicate that these representations might be incomplete for other tasks and healthcare applications.

Now, all the DyNeMoC models outperformed all baselines models across all subjects, except for Subject 3 (see Table 4) where all models performed poorly (which might indicate a data collection error or protocol issue for that individual). Moreover, the DyNeMoC-Transformer model achieved far superior test accuracies than the DyNeMoC-RNN-Large model which had a greater number of parameters. This is expected as the transformer variant processed trials as a whole rather than one time step at a time like the RNN one did and utilized the self-attention mechanism (Vaswani et al., 2017) for finding relationships between timesteps which the RNN one did not.

Overall, our results demonstrate that our proposed semi-supervised approach — DyNeMoC, in general, and DyNeMoC-Transformer, in particular – is a simple, robust, and very effective approach for learning useful latent representations from noisy data for downstream classification and potentially other healthcare applications.

## 5 Conclusion

In this work, we investigated an RNN-VAE model named DyNeMo which has been designed to model dynamic brain networks. Specifically, we evaluated the utility of the latent brain network description provided by DyNeMo in a downstream classification task. We found that the latent representations obtained from the unsupervised model alone were not sufficient to properly perform the task, which may also apply to other healthcare-related tasks of interest. We, therefore, proposed a semi-supervised architecture named DyNeMoC that jointly trained DyNeMo and an MLP classifier to optimize for both the variational free energy and cross-entropy. We showed this was crucial for improving the performance of the classification. We further demonstrated that the transformer variant of DyNeMoC outperformed the RNN one. Finally, we note that we focused on improving individual classification accuracies because, for healthcare applications, such as personalized treatment of neurodegenerative diseases like Alzheimer's and Parkinson's and the construction of customized brain-computer interfaces, we are interested in making individualized predictions. We believe that we can leverage information across individuals by training models with multiple subjects and fine-tuning. This could also potentially help improve the individual subject predictions. We leave this for future work.

## REFERENCES

Elena A. Allen, Eswar Damaraju, Sergey M. Plis, Erik B. Erhardt, Tom Eichele, and Vince D. Calhoun. Tracking Whole-Brain Connectivity Dynamics in the Resting State. *Cerebral Cortex*, 24(3):663–676, 11 2012. ISSN 1047-3211. doi: 10.1093/cercor/bhs352. URL https://doi.org/10.1093/cercor/bhs352.

Claudio Babiloni, Katarzyna Blinowska, Laura Bonanni, Andrej Cichocki, Willem De Haan, Claudio Del Percio, Bruno Dubois, Javier Escudero, Alberto Fernández, Giovanni Frisoni, et al. What electrophysiology tells us about alzheimer's disease: a window into the synchronization and connectivity of brain neurons. *Neurobiology of aging*, 85:58–73, 2020.

A.P. Baker, M.J. Brookes, I.A. Rezek, S.M. Smith, T. Behrens, P.J.P. Smith, and M. Woolrich. Fast transient networks in spontaneous human brain activity. *eLife*, 2014(3), 2014. doi: 10.7554/eLife.01867.

Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Steven L Bressler and Vinod Menon. Large-scale brain networks in cognition: emerging methods and principles. *Trends in cognitive sciences*, 14(6):277–290, 2010.

György Buzsáki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304 (5679):1926–1929, 2004.

Thomas A Carlson, Hinze Hogendoorn, Ryota Kanai, Juraj Mesik, and Jeremy Turret. High temporal resolution decoding of object position and category. *Journal of vision*, 11(10):9–9, 2011.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28:2980–2988, 2015.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Similarity-based fusion of meg and fmri reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, 26(8):3563–3579, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Stavros I Dimitriadis, María E López, Ricardo Bruña, Pablo Cuesta, Alberto Marcos, Fernando Maestú, and Ernesto Pereda. How to build a functional connectomic biomarker for mild cognitive impairment from source reconstructed meg resting-state activity: the combination of roi representation and connectivity estimator matters. *Frontiers in neuroscience*, 12:306, 2018.

Elise G.P. Dopper, Serge A.R.B. Rombouts, Lize C. Jiskoot, Tom den Heijer, J. Roos A. de Graaf, Inge de Koning, Anke R. Hammerschlag, Harro Seelaar, William W. Seeley, Ilya M. Veer, Mark A. van Buchem, Patrizia Rizzu, and John C. van Swieten. Structural and functional brain connectivity in presymptomatic familial frontotemporal dementia. *Neurology*, 83(2): e19–e26, 2014. ISSN 0028-3878. doi: 10.1212/WNL.0000000000000583. URL https://n.neurology.org/content/83/2/e19.

Otto Fabius and Joost R Van Amersfoort. Variational recurrent auto-encoders. *arXiv preprint arXiv:1412.6581*, 2014.

Eleonora Fiorenzato, Antonio P Strafella, Jinhee Kim, Roberta Schifano, Luca Weis, Angelo Antonini, and Roberta Biundo. Dynamic functional connectivity changes associated with dementia in Parkinson's disease. *Brain*, 142(9):2860–2872, 07 2019. ISSN 0006-8950. doi: 10.1093/brain/awz192. URL https://doi.org/10.1093/brain/awz192.

Michael D Fox and Marcus E Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience*, 8(9):700–711, 2007.

Karl J. Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2(1-2):56–78, 1994. doi: https://doi.org/10.1002/hbm.460020107. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.460020107.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

Chetan Gohil, Evan Roberts, Ryan Timms, Alex Skates, Cameron Higgins, Andrew Quinn, Usama Pervaiz, Joost van Amersfoort, Pascal Notin, Yarin Gal, Stanislaw Adaszewski, and Mark Woolrich. Mixtures of large-scale dynamic functional brain network modes. *NeuroImage*, 263:119595, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119595. URL https://www.sciencedirect.com/science/article/pii/S1053811922007108.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL http://arxiv.org/abs/1606.08415.

R. Hindriks, M.H. Adhikari, Y. Murayama, M. Ganzetti, D. Mantini, N.K. Logothetis, and G. Deco. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fmri? *NeuroImage*, 127:242–256, 2016. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.11.055. URL https://www.sciencedirect.com/science/article/pii/S1053811915010782.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Leyla Isik, Ethan M Meyers, Joel Z Leibo, and Tomaso Poggio. The dynamics of invariant object recognition in the human visual system. *Journal of neurophysiology*, 111(1):91–102, 2014.

Sanja Josef Golubic, Cheryl J Aine, Julia M Stephen, John C Adair, Janice E Knoefel, and Selma Supek. Meg biomarker of alzheimer's disease: Absence of a prefrontal generator during auditory sensory gating. *Human Brain Mapping*, 38(10):5180–5194, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

Zeb Kurth-Nelson, Gareth Barnes, Dino Sejdinovic, Ray Dolan, and Peter Dayan. Temporal structure in associative retrieval. *Elife*, 4:e04919, 2015.

Giulia Liberati, Josué Luiz Dalboni Da Rocha, Linda Van der Heiden, Antonino Raffone, Niels Birbaumer, Marta Olivetti Belardinelli, and Ranganatha Sitaram. Toward a brain-computer interface for alzheimer's disease patients by combining classical conditioning and brain state classification. *Journal of Alzheimer's Disease*, 31(s3):S211–S220, 2012.

Fernando Lopes da Silva. Eeg and meg: Relevance to neuroscience. *Neuron*, 80(5):1112–1128, 2013. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2013.10.017. URL https://www.sciencedirect.com/science/article/pii/S0896627313009203.

Pravat K Mandal, Anwesha Banerjee, Manjari Tripathi, and Ankita Sharma. A comprehensive review of magnetoencephalography (meg) studies for brain functionality in healthy aging and alzheimer's disease (ad). *Frontiers in computational neuroscience*, 12:60, 2018.

Shiv Kumar Mudgal, Suresh K Sharma, Jitender Chaturvedi, and Anil Sharma. Brain computer interface advancement in neurosciences: Applications and issues. *Interdisciplinary Neurosurgery*, 20:100694, 2020.

William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

Yonatan Sanz Perl, Hernán Bocaccio, Ignacio Pérez-Ipiña, Federico Zamberlán, Juan Piccinini, Helmut Laufs, Morten Kringelbach, Gustavo Deco, and Enzo Tagliazucchi. Generative embeddings of brain collective dynamics using variational autoencoders. *Phys. Rev. Lett.*, 125:238101, Dec 2020. doi: 10.1103/PhysRevLett.125.238101. URL `https://link.aps.org/doi/10.1103/PhysRevLett.125.238101`.

Malcolm Proudfoot, Mark W Woolrich, Anna C Nobre, and Martin R Turner. Magnetoencephalography. *Practical Neurology*, 14(5):336–343, 2014. ISSN 1474-7758. doi: 10.1136/practneurol-2013-000768. URL `https://pn.bmj.com/content/14/5/336`.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Marcus E Raichle, Ann Mary MacLeod, Abraham Z Snyder, William J Powers, Debra A Gusnard, and Gordon L Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001.

Deborah N Schoonhoven, Matteo Fraschini, Prejaas Tewarie, Bernard MJ Uitdehaag, Anand JC Eijlers, Jeroen JG Geurts, Arjan Hillebrand, Menno M Schoonheim, Cornelis J Stam, and Eva MM Strijbis. Resting-state meg measurement of functional activation as a biomarker for cognitive decline in ms. *Multiple Sclerosis Journal*, 25(14):1896–1906, 2019.

Julia Schumacher, Luis R Peraza, Michael Firbank, Alan J Thomas, Marcus Kaiser, Peter Gallagher, John T O'Brien, Andrew M Blamire, and John-Paul Taylor. Dysfunctional brain dynamics and their origin in Lewy body dementia. *Brain*, 142(6):1767–1782, 04 2019. ISSN 0006-8950. doi: 10.1093/brain/awz069. URL `https://doi.org/10.1093/brain/awz069`.

Yoshihito Shigihara, Hideyuki Hoshi, Jesús Poza, Víctor Rodríguez-González, Carlos Gómez, and Takao Kanzawa. Predicting the outcome of non-pharmacological treatment for patients with dementia-related mild cognitive impairment. *Aging (Albany NY)*, 12(23):24101, 2020a.

Yoshihito Shigihara, Hideyuki Hoshi, Keita Shinada, Toyoji Okada, and Hajime Kamada. Non-pharmacological treatment changes brain activity in patients with dementia. *Scientific reports*, 10 (1):1–9, 2020b.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Diego Vidaurre, Andrew J. Quinn, Adam P. Baker, David Dupret, Alvaro Tejero-Cantero, and Mark W. Woolrich. Spectrally resolved fast transient brain states in electrophysiological data. *NeuroImage*, 126:81–95, 2016. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.11.047. URL `https://www.sciencedirect.com/science/article/pii/S1053811915010691`.

Diego Vidaurre, Stephen M. Smith, and Mark W. Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1705120114. URL `https://www.pnas.org/content/114/48/12827`.

Diego Vidaurre, Romesh Abeysuriya, Robert Becker, Andrew J. Quinn, Fidel Alfaro-Almagro, Stephen M. Smith, and Mark W. Woolrich. Discovering dynamic brain networks from big data in rest and task. *NeuroImage*, 180:646–656, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.06.077. URL `https://www.sciencedirect.com/science/article/pii/S1053811917305487`. Brain Connectivity Dynamics.

Fabrice Wendling, Karim Ansari-Asl, Fabrice Bartolomei, and Lotfi Senhadji. From eeg signals to brain connectivity: A model-based evaluation of interdependence measures. *Journal of Neuroscience Methods*, 183(1):9–18, 2009. ISSN 0165-0270. doi: https://doi.org/10.1016/j.jneumeth.2009.04.021. URL `https://www.sciencedirect.com/science/article/pii/S0165027009002350`. BrainModes: A Principled Approach to Modeling and Measuring Large-Scale Neuronal Activity.

Qiong Zhang, Jelmer P Borst, Robert E Kass, and John R Anderson. Inter-subject alignment of meg datasets in a common representational space. Technical report, Wiley Online Library, 2017.

# A Appendix

## A.1 Prior Work

The usage of sliding windows used to be the predominant technique for studying neural dynamics. This technique (and other methods that build upon it) is, however, limited by the necessity of manually specifying the temporal window, i.e., the time scale at which the neural activities of interest take place (Hindriks et al., 2016). This manual specification usually needs to deal with the critical trade-off in two conflicting criteria: the time window being too long leads to missing fast dynamics and the time window being too short leads to insufficient data for making reliable network estimation.

An alternative approach that overcomes the shortcomings of sliding windows is the usage of generative models such as HMM-based models which can describe neural activity as a dynamic sequence of discrete brain states where each state is characterized by distinct network activity patterns. An HMM can be trained in an unsupervised way, and the learned state sequence of the HMM can be connected to task timings post-hoc to reveal task-induced neural dynamics (Vidaurre et al., 2016). However, HMM-based models are limited by the Markov assumption that the activation of states at a particular time point only depends on the activation of states at the previous time point. This disregards the rich dynamic courses that states undergo to arrive at a particular probabilistic configuration at a given moment of time.

As described in Gohil et al. (2022), DyNeMo was designed to overcome the shortcomings of HMM-based models.

## A.2 Model Architectures and Training

Table 2: Architectures of DyNeMoC Models

| | **DyNeMoC RNN-Small** | **DyNeMoC RNN-Large** | **DyNeMoC Transformer** |
|---|---|---|---|
| Encoder - Network Type | LSTM | LSTM | BERT |
| Encoder - Hidden Size | 128 | 416 | 128 |
| Encoder - Layers | 1 | 2 | 1 |
| Encoder - Attention Heads | | | 1 |
| Prior Network - Network Type | LSTM | LSTM | GPT2 |
| Prior Network - Hidden Size | 128 | 416 | 64 |
| Prior Network - Layers | 1 | 2 | 1 |
| Prior Network - Attention Heads | | | 1 |
| MLP - Hidden Size | 1024 | 1024 | 1024 |
| MLP - Layers | 2 | 2 | 2 |
| MLP - Dropout | 0.5 | 0.5 | 0.9 |
| Total # of Parameters | 2.1M | 9.8M | 9.2M |

In all our experiments, we fixed the number of latent states to 20 and trained all the models three times for 200 epochs (of which the first 100 had $\tanh$ KL annealing (Fu et al., 2019)). Moreover, we used the Adam (Kingma & Ba, 2014) optimizer with a learning rate of 1e-3 and set the batch size to 64. The rest of the hyperparameters were the same as in Gohil et al. (2022).

### A.3 CHOOSING THE VALUE OF CROSS-ENTROPY COEFFICIENT

To select the appropriate value for the cross-entropy coefficient $w_c$, we trained DyNeMoC-RNN-Small on the 1st subject with $w_c$ ranging between 1 to $10^6$. As shown in Table 3, the validation accuracy became competitive from $w_c = 10^3$ onward and was the highest at $w_c = 10^4$. This made sense as the free energy term in the loss function of DyNeMoC was also in the order of $10^4$.

Table 3: Classification accuracies of DyNeMoC-RNN-Small for different values of $w_c$ on the validation set of the 1st subject

| $\boldsymbol{w_c}$ | $\mathbf{10^0}$ | $\mathbf{10^1}$ | $\mathbf{10^2}$ | $\mathbf{10^3}$ | $\mathbf{10^4}$ | $\mathbf{10^5}$ | $\mathbf{10^6}$ |
|---|---|---|---|---|---|---|---|
| **Accuracy** | 0.030 | 0.041 | 0.068 | 0.311 | **0.379** | 0.351 | 0.324 |

Hence, we set $w_c$ to $10^4$ in all of our experiments.

### A.4 ADDITIONAL RESULTS

Table 4: Classification accuracies on the test sets of different subjects

| **Subject** | **SVM Baseline** | **MLP Baseline** | **DyNeMo RNN-Small + MLP** | **DyNeMo RNN-Large + MLP** | **DyNeMo Transformer + MLP** | **DyNeMoC RNN-Small** | **DyNeMoC RNN-Large** | **DyNeMoC Transformer** |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.257 | 0.261 | 0.060 | 0.062 | 0.075 | 0.359 | 0.368 | **0.520** |
| 2 | 0.136 | 0.116 | 0.078 | 0.037 | 0.063 | 0.186 | 0.187 | **0.273** |
| 3 | **0.066** | 0.061 | 0.030 | 0.016 | 0.027 | 0.049 | 0.064 | 0.056 |
| 4 | 0.311 | 0.326 | 0.112 | 0.049 | 0.065 | 0.520 | 0.504 | **0.668** |
| 5 | 0.521 | 0.576 | 0.230 | 0.156 | 0.139 | 0.687 | 0.691 | **0.820** |
| 6 | 0.248 | 0.288 | 0.091 | 0.067 | 0.094 | 0.388 | 0.413 | **0.568** |
| 7 | 0.307 | 0.334 | 0.165 | 0.124 | 0.137 | 0.498 | 0.521 | **0.668** |
| 8 | 0.121 | 0.154 | 0.032 | 0.063 | 0.034 | 0.145 | 0.158 | **0.192** |
| 9 | 0.275 | 0.294 | 0.068 | 0.046 | 0.046 | 0.415 | 0.415 | **0.559** |
| 10 | 0.182 | 0.234 | 0.121 | 0.101 | 0.044 | 0.324 | 0.308 | **0.438** |
| 11 | 0.117 | 0.163 | 0.024 | 0.028 | 0.028 | 0.168 | 0.210 | **0.274** |
| 12 | 0.074 | 0.080 | 0.016 | 0.012 | 0.004 | 0.087 | 0.095 | **0.117** |
| 13 | 0.036 | 0.042 | 0.013 | 0.018 | 0.009 | 0.051 | 0.049 | **0.057** |
| 14 | 0.184 | 0.177 | 0.048 | 0.018 | 0.046 | 0.313 | 0.352 | **0.458** |
| 15 | 0.133 | 0.110 | 0.023 | 0.038 | 0.037 | 0.169 | 0.180 | **0.256** |
| **Mean (Std. Err.)** | 0.20 (0.04) | 0.21 (0.05) | 0.07 (0.02) | 0.06 (0.02) | 0.06 (0.01) | 0.29 (0.07) | 0.30 (0.07) | **0.40 (0.09)** |