



**SAPIENZA**  
UNIVERSITÀ DI ROMA

# **A novel transformer-based approach for estimating causal interaction in multichannel electroencephalographic data**

**Facoltà di Ingegneria dell'informazione, informatica e statistica  
Dipartimento di Ingegneria informatica, automatica e gestionale  
Corso di Laurea Magistrale in Artificial Intelligence and Robotics**

**Luca Maurici**  
**Matricola 1809678**

Relatore  
Prof.ssa Laura Astolfi

Correlatore  
Prof. Nicola Toschi

A.A. 2021-2022



# Summary

<b>INTRODUCTION .....</b>	<b>1</b>
OVERVIEW AND STATE OF THE ART .....	1
PREMISES .....	5
AIMS .....	6
OUTLINE.....	8
<b>1. BASIC OF CONNECTIVITY IN NEUROSCIENCE .....</b>	<b>9</b>
1.1 PHYSIOLOGY OF THE NERVOUS SYSTEM .....	9
1.2 ELECTROENCEPHALOGRAPHY.....	12
1.3 BRAIN CONNECTIVITY .....	16
1.3.1 Anatomical and functional connectivity.....	17
1.3.2 Connectivity graphs.....	17
<b>2. TRADITIONAL METHODS FOR ESTIMATING WIENER-GRANGER CAUSALITY .....</b>	<b>21</b>
<b>3. REAL AND SIMULATED ENCEPHALOGRAPHIC DATA .....</b>	<b>26</b>
3.1 SYNTHETIC EEG DATA .....	26
3.2 REAL HYPERSCANNING EEG DATA.....	29
3.2.1 Experiment description.....	30
3.2.2 Data processing and structure .....	32
<b>4. PROPOSED METHODS .....</b>	<b>34</b>
4.1 TRANSFORMER MODELS.....	34
4.1.1 Input and output .....	35
4.1.2 Model architecture.....	35
4.1.3 Multi-head attention mechanisms.....	37
4.1.4 Position-wise feed-forward networks .....	38
4.1.5 Positional Encoding .....	39
4.2 THE SPACETIMEFORMER ARCHITECTURE .....	39
4.2.1 Input sequence .....	41
4.2.2 Spatiotemporal embedding.....	42
4.3 EXPERIMENTS .....	45
4.3.1 Reference method for CGC estimation applied to synthetic data .....	46
4.3.2 Novel method for CGC estimation applied to synthetic data .....	48
4.3.3 Novel attention-based method for causality estimation applied to hyperscanning data .....	53

**5. RESULTS ..... 59**

5.1 STATISTICAL TEST ON THE NOVEL METHOD FOR CGC ESTIMATION APPLIED TO SYNTHETIC DATA ..... 59

5.2 STATISTICAL TEST ON THE NOVEL ATTENTION-BASED METHOD FOR CAUSALITY ESTIMATION APPLIED TO  
HYPERSCANNING DATA ..... 63

**CONCLUSION AND FUTURE WORKS ..... 69**

**REFERENCES..... 72**

# Introduction

## Overview and State of the Art

This thesis is based on the idea of applying artificial intelligence methods to neuroscience, two fields of research that, by dealing with intelligent systems, have similar groundings in some respects. There is, however, at present, a relatively modest number of studies that attempt to solve neuroscience problems using artificial intelligence methods.

Neuroscience is a field of research with boundless potential for what is still discoverable, yet at the same time, it represents a challenge for the scientific community due to problems such as the extreme complexity of intelligent systems, limitations of measurement instrumentation, and ethical issues.

The design of a scientific experiment in this field begins precisely with the choice of instrumentation, which must be a compromise between invasiveness and accuracy of measurements, depending on the type and needs of the study.

Due to its zero invasiveness, excellent temporal resolution and capability to capture the oscillatory nature of the activity of large neuronal ensembles, one of the most popular recording approaches in neuroscience is electroencephalography (EEG), which allows, through the application of electrodes on the scalp, to record the electrical activity of the brain during a time span. The measurements are subject to a considerable

amount of noise and artifacts, which must be mitigated by the reiteration of the experiment task for a considerable number of times and, at a later stage, through signal processing and appropriate data modelling. The approach I propose in this thesis is based on electroencephalographic data of both synthetic and real nature.

In order to study the data collected from an electroencephalographic experiment, it is often useful, if not necessary, to estimate a network representing the functional connectivity of the brain of the subject undergoing the experiment. The connectivity network, that can be seen as a graph, can be employed directly, without further processing, to make classifications or regressions; or used as input to algorithms that calculate numerical indices that quantify certain properties of the graph, on which it is common to make correlations with subject behavioural data, clustering, statistical analysis, classification, and regression.

This thesis focuses on the problem of estimating the connectivity network from EEG data.

Traditional state-of-the-art approaches compute cerebral connectivity interpreted as correlation (e.g., Ordinary Coherence (Kaminski & Blinowska, 2022)) or cause-effect relationship (e.g., Conditional Granger causality (Seth, 2007), Partial Directed Coherence (Baccala & Sameshima, 2001)) between different encephalographic signals from electrodes placed on the subject's scalp.

Ordinary Coherence (Kaminski & Blinowska, 2022) is a spectral method which computes the linear correlation between two signals at a given

frequency. There are, alternatively, some methods capable of modelling the directionality of interaction and which are based on the concept of Granger causality (Seth, 2007) that is based on the reduction of the residuals of an autoregressive predictive model, seen as a sign of causality in the statistical sense (Wiener, et al., 1949). There are some variants of the same basic idea, such as the Bivariate and Multivariate (or Conditional) Granger Causality Test (Seth, 2007) which provides an estimate also of the directionality of the interaction and are based on Linear Autoregressive Models (Schlögl, 2006). Partial Directed Coherence (PDC) (Baccala & Sameshima, 2001) is a spectral extension of these methods and one of the most widely used. Such multichannel versions of the test are based on a Multivariate Autoregressive (MVAR) (Schlögl, 2006) modeling.

Recent deep learning (DL) approaches (Shrestha & Mahmood, 2019) have proved to be very effective in helping to derive meaning from EEG signals thanks to their ability to learn good feature representations from raw data (Roy, et al., 2019).

In addition, Artificial Neural Networks (Faes, Vantieghem, & Van Hulle, 2022) are recognised as an extremely interesting model, given their degree of flexibility in approximating strongly non-linear relationships between variables and the fact that no *a priori* assumptions need to be made about signal stationarity and connectivity pattern. Faes et al. (Faes, Vantieghem, & Van Hulle, 2022), for instance, concluded that some kinds of neural networks, in particular LSTMs (Yu, Si, Hu, & Zhang, 2019), can be employed to estimate the direct connectivity of reconstructed EEG sources.

The framework developed by (Ying, Bourgeois, You, Zitnik, & Leskovec, 2019) for graph neural networks can be used to carry out some sort of connectivity estimation. After training a GNN model on a classification task, the GNNexplainer (Ying, Bourgeois, You, Zitnik, & Leskovec, 2019) framework is used to determine some features and properties of the analysed graphs, which are crucial in training GNN: for example, deriving the importance of nodes and edges in the network.

Although deep learning approaches are showing great potential and good performances in different fields, there is still a limited literature on the use of deep learning approaches to estimate brain causality.

Transformer architectures were introduced by Google Brain in the well-known paper “Attention is all you need” (Vaswani, et al., 2017), initially finding application in the field of natural language processing (NLP) (Nadkarni, Ohno-Machado, & Chapman, 2011). Various variants adapted to diverse types of problems are becoming increasingly used in different fields (e.g., computer vision, time series forecasting), as they are proving to be extremely performant, such that they are becoming the state of the art in many applications. Transformers, such as Recurrent Neural Networks (RNNs) (Yu, Si, Hu, & Zhang, 2019), are designed to process sequential data in the same manner as natural language, for tasks such as text summarisation and translation. However, as opposed to RNNs, Transformers do not require to process sequential data in order. For example, if the input data is a natural language sentence, then a transformer does not need to process the beginning before the end. This



property allows the transformers to parallelise much more than an RNN, thus reducing the training time.

## Premises

The aforementioned properties of transformer architectures, combined with their ability to deal with time series, makes transformer architectures excellent candidates for processing electroencephalographic data.

In particular, amongst all the different implementations of transformer-inspired models for multivariate forecasting, I focused on the Spacetimeformer (Grigsby, Wang, & Qi, 2021). Its main peculiarity is the ability to compute both temporal and spatial attention, avoiding indirect information flow through the network, allowing the model to directly pay attention to the most relevant parts of input information, along the two dimensions of temporal lags and channels.

In the brain, non-linearities are present at the level of the single neuron in the propagation of the electrochemical signal. This is reflected also at the macroscopic level, with non-linear relations between electroencephalographic signals (i.e., electrodes). This is the reason why it is interesting to use non-linear models to express the aforementioned relationships, as opposed to traditional linear methods currently used in neuroscience (e.g., Conditioned Granger Causality, PDC, etc.).

Another important advantage of using non-linear machine learning methods lies in the fact that they do not impose constraints on the model,

based on a priori physiological assumptions, as is the case for some existing non-linear models.

Finally, hyperscanning (Babiloni & Astolfi, 2014) is a technique that allows for the simultaneous analysis of the brain activity of two or more subjects as they interact during a social task. Hyperscanning data represents a great opportunity to test non-linear models, because the nature of inter-subject causal dependences is still not described in the literature, and it probably implies non-linear relationships that current linear models may not be able to approximate at an acceptable level, as they currently do for intra-subject causal relationships.

## Aims

The general scope of this thesis is to study the applicability of two novel transformer-based methods to estimate causality on EEG data. Given all the premises made, the two approaches, with their specific aims, can be summarized into two main points:

- The first aim is to use the Spacetimeformer model as a non-linear predictive model to replace MVAR linear models in Conditional Granger Causality method.

The goal is to assess the feasibility of the novel approach, using synthetic EEG data as a benchmark to compare results with a ground truth. The hypothesis is that the novel approach can show performances comparable to those of the Linear Conditioned

Granger causality method. The application of the method on real data is dependent on computational times, which are expected to be a possible criticality.

- The second aim is to leverage the explainability properties of the Spacetimeformer architecture to estimate brain connectivity by inspecting computed attention. My hypothesis is that the attention computed by the model may be seen as a form of causality in the predictive sense, and a combination of them may be used as the weights of the connectivity graph to be estimated. In this case, computational times are expected to be sensibly lower with respect to the first method. For this reason, the aim is to assess the plausibility of this novel method by applying it to real hyperscanning EEG data. The choice of this kind of data is motivated by, for two reasons:

While it is known, in literature, that causal relationships in the brain can be modelled linearly without heavily affecting the causality estimation, if inter-subject causal relationship (i.e., the ones between the two brains) can be approximated to linear dependences is still an open question. The proposed method can therefore prove to be an effective tool able to capture the intrinsic nature of such relationship, without the need of any a priori assumptions.

The nature of multi-subject networks associated with hyperscanning techniques provides a straightforward interpretation of the main network properties, that can be used to validate the method.

Finally, to the best of my knowledge, this is the first attempt to apply a Deep Learning approach to hyperscanning data.

## Outline

Chapter 1 consists of a general background on nervous system physiology, electroencephalography, and hints of connectivity graphs.

Chapter 2 discusses state-of-the-art methods for estimating Weiner-Granger causality, which is later recalled in Chapter 4, which discusses the two novel methods I proposed as well as the one taken as a comparison. Chapter 4 also discusses the experiments performed.

In Chapter 3, I illustrate the data used in the experiments, both synthetically generated pseudo-EEG data and real hyperscanning EEG data.

Finally, in Chapter 5 the results of the experiments are discussed.

The Conclusion summarizes the main findings of this thesis and outlines the possible avenues for future development of the research here introduced.

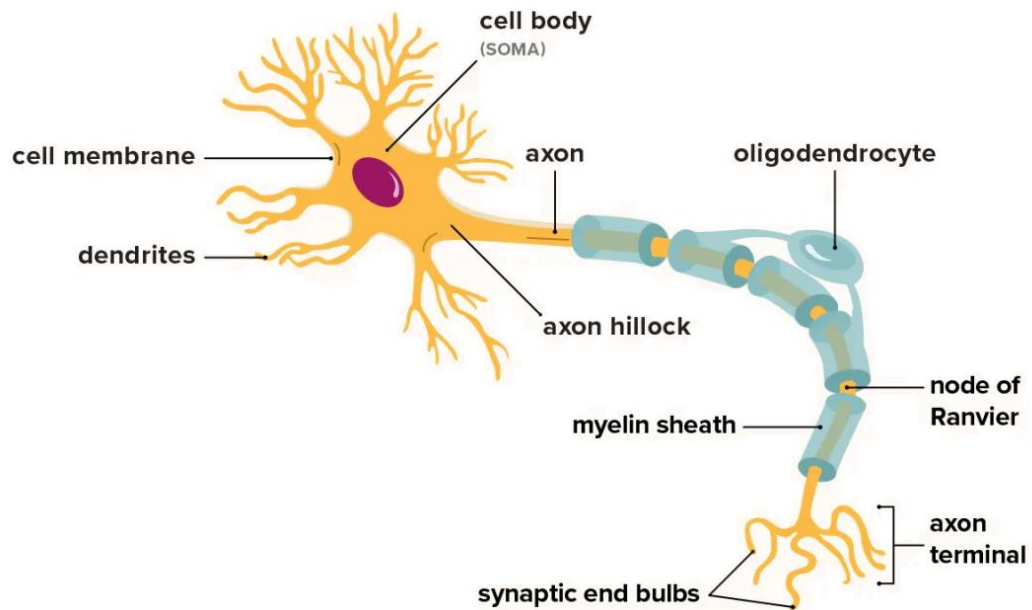
# Chapter 1

## Basic of connectivity in neuroscience

### 1.1 Physiology of the nervous system

The nervous system is responsible for the coordination of voluntary and involuntary motor and cognitive activities. It can be divided into the peripheral nervous system, which can be further subdivided into several components, each of which has a distinct purpose, and the central nervous system.

The spinal cord and the brain constitute the central nervous system. The latter, contained in the cranial box, continues directly into the former. The central nervous system takes in information, processes it, organises internal and external stimuli and makes the necessary plans. Glial cells and nerve cells, or neurons, constitute most of the cells in the central nervous system. The former play supporting roles helping to create the brain barrier and myelin. The latter constitute the basic anatomical element of the neurological system.



*Figure 1: Anatomy of a neural cell. From (Smith, 2022).*

The basic functions of a neuron (Figure 1) are collecting information from various sources; the integration function: processing incoming information (in time and space to provide a binary decision); and generating and propagating the bit of information (binary decision) to the target cells (e.g., other neural cells, muscle cells). In essence, the cellular body, by collecting information from other neurons, can make a single binary decision, which is then propagated to all other cells to which it is output connected.

The cell body, which houses the nucleus, and the extensions--the axon on one side and the dendrites on the other--constitute a neuron. The cell body is enclosed in a thin protective membrane called the plasma membrane.

While the axon is a fibre that enables signal transmission, dendrites are the main sites for receiving signals from neighbouring neurons.

Neurons, which are excitable cells, have a resting membrane potential that lies between -60 mV and -70 mV.

The electrochemical equilibrium makes the resting potential to converge to this value, which can be obtained through the Nernst equation:

$$\Delta\mu = RT \ln \frac{[X]_A}{[X]_B} + zF(E_A - E_B)$$

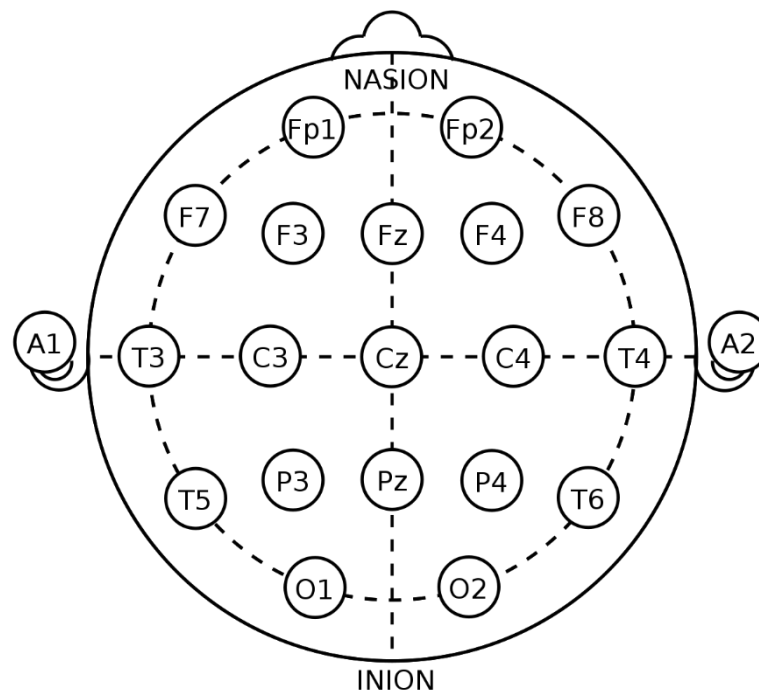
Where the first term represents the diffusional forces produced by the difference in concentration of ions in the intra- and extracellular fluid; while the second term represents the electrical forces caused by the attraction of positive ions to regions of negative potential. The neuron membrane has some basic structures called ion pumps, which control the traversal of the membrane by ions.

Under certain circumstances, membrane-crossing ionic currents can generate changes in the membrane potential that, if a certain potential value is reached, cause the so-called action potential.

The effect of this change spreads to other cells carrying information. Such a process is binarized by a threshold: if the stimulus does not reach a certain threshold, it does not occur. If, in contrast, the threshold is reached, it will always have the same form, duration and intensity, independently of the amplitude of the stimulus. Action potentials represent the binary decision of neurons, which starting in one cell, they propagate to many others (Avanzolini & Magosso, 2015).

## 1.2 Electroencephalography

Brain activity can be measured both intracellularly and extracellularly. Electroencephalography (EEG) is a non-invasive technique for measuring extracellular electrical activity of the brain by applying a series of electrodes to the scalp, useful for evaluating the function of the central nervous system. The EEG signal, for each electrode, is obtained as the difference in measured potentials by the electrode of interest and a reference one.



*Figure 2: 10-20 system electrodes' placement.*

The most influential nerve structure contributing to the generation of EEG signals is the cerebral cortex. It is composed of six layers, consisting of two types of neurons: non-pyramidal and pyramidal. The former are characterised by the fact of having very short dendrites that grow in all directions, allowing connections between adjacent neurons; the



pyramidal, on the other hand, are so called due to the shape of their soma, representing about 3/4 of all the cortical neurons. They have an apical dendrite that allows them to communicate with more superficial neurons. The dendrites are arranged in parallels to each other and perpendicular to the cortex surface. Hence, pyramidal neurons contribute mostly to the generation of the electroencephalographic signal with their postsynaptic activity, given the peculiar arrangement of their dendrites.

Nevertheless, the electrical signal detected by the electrodes has low voltage, so the acquisition system must comprise several stages, including amplification. The major problems with encephalographic data are the low amplitude of the signals, jointly with an important level of susceptibility to artefacts (e.g., muscles' contraction, eye movements), as well as an intrinsically compromised spatial resolution. The track recorded by EEG represents the graphical recording over time of the electrical activity generated by communities of thousands of neurons, composed by a succession of waves with different frequencies and amplitudes.

Thanks to his fine temporal resolution, EEG can be employed in different applications including:

1. detecting signals in real time allowing BCI (Nicolas-Alonso & Gomez-Gil, 2012) use;
2. studying and visualizing the amount of brain activity in the different areas;
3. and diagnosing brain disorders.

The key differences between various EEG acquisition systems are the electrodes utilized, their quantity, sampling frequency, and other factors.

Although the majority of information is contained within a frequency range of 40 Hz, the EEG signal's bandwidth is between 1 and 50 Hz. The different brain rhythms that constitute the EEG signal can be distinguished based on the different oscillation frequencies of the waves of neuronal activity.

In general, different brain rhythms are associated with different cognitive tasks, reflecting the synchronous and coherent activity of the participating populations of neurons (Klimesch, 1999).

The five brain rhythms are the following.

- Gamma rhythm: frequencies above 30 Hz, but low amplitudes of around 15  $\mu\text{V}$ . The gamma rhythm occurs in the presence of high cognitive functionality, including processes that allow the integration of information from different cortical areas.
- Beta rhythm: average voltage of about 19  $\mu\text{V}$  and frequencies between 13 and 30 Hz. The frontal regions of the cortex are good places to look for it. It is peculiar of states of attention, focus, and alertness and contributes to the coordination of motor functions. When there are numerous environmental stimuli, beta rhythm is detected.
- Alpha rhythm: frequency range of 8 to 12 Hz. It has a voltage of about 40  $\mu\text{V}$  on average. It has a similar sinusoidal rhythm and is

more prevalent in the occipital regions of the cortex. When the mind is relaxed, when the eyes are closed while awake, and when there are few external stimuli present, this rhythm is normal.

- Theta rhythm: frequencies between 4 and 7 Hz are its defining characteristics. It contributes to the encoding and retrieval of episodic memories as well as deep sleep states. In visual imagery and hypnopompic imagery.
- Delta rhythm: frequencies between 1 and 3 Hz and amplitudes up to 200  $\mu$ V. Although it can also be heard in wakefulness or hazy dream states, this rhythm is usually associated with deep sleep conditions.

EEG signals are particularly sensitive to noise and artifacts, as one might expect. These distortions can come from a variety of sources, including improperly applied electrodes, electrode small movements, and alternating current-related artefacts. The most common sources of artefacts caused by the human body are eye movements and blinks, muscle contractions, sweating, and cardiac activity. Therefore, a robust pre-processing approach aiming at getting a much cleaner signal is a crucial step before using and interpreting EEG signals (He, Yuan, Meng, & Gao, 2020).

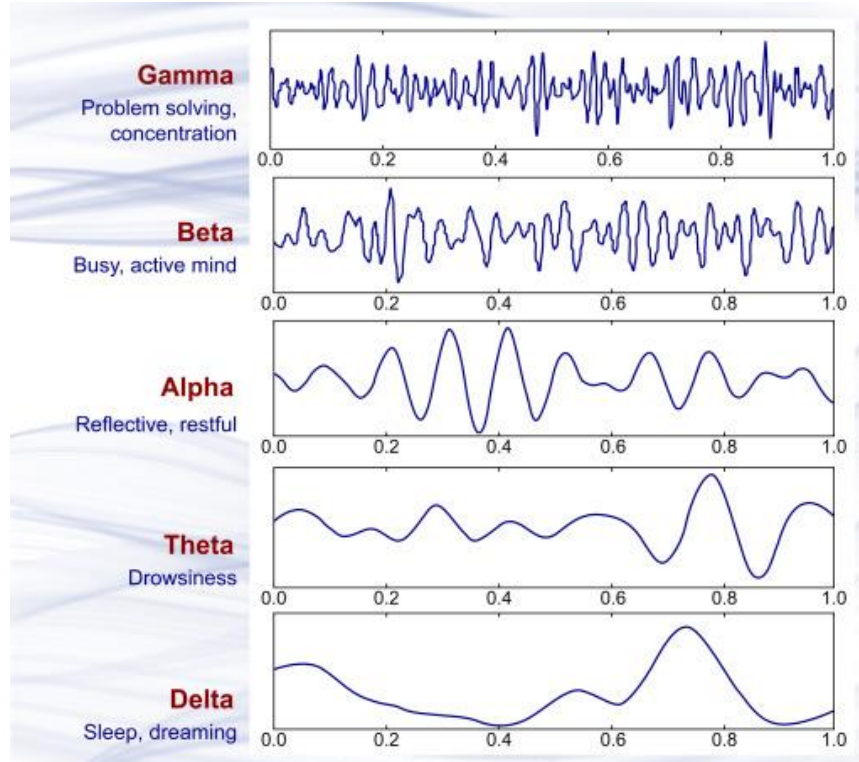


Figure 3: Different brain rhythms are distinguished from each other by the different oscillation frequency of neuronal activity waves. Each rhythm is associated to some brain functions. From (Abhang, Gawali, & Mehrotra, 2016).

## 1.3 Brain connectivity

The brain is a complex system that contains dynamic brain networks at many scales, including neurons, brain regions, and social systems. Network neuroscience, which examines the brain from this angle, tries to record, map, and model the components and interactions between neurological systems.

### 1.3.1 Anatomical and functional connectivity

Brain networks are distinguished by two main distinct connection natures.

1. Anatomical connectivity, which explains the anatomical or physical links tying together groups of neurons or other neural components. It is not expected to be altered over the course of a few-hour recording session because it is very stable at short time scales. Functional and effective connectivity are based on anatomical connectedness, although this connectivity cannot be explained.
2. Functional connectivity is a way to describe what is going on in the brain in terms of how comparable the activity in certain regions is to one another. Typically, it is grounded on transfer entropy, coherence, correlation, and causality estimation. In particular, the latter aims to track the effect of one neurological system has over another, either at the synaptic or population level.

### 1.3.2 Connectivity graphs

The most suitable way to represent a network in general, and in this case a brain network is using a graph. It can be formally defined as a tuple  $G = (V, E)$  where  $V$  is the set of the  $N = |V|$  nodes and  $E \subseteq V \times V$  is the set of the  $M = |E|$  edges linking the nodes. It is possible to represent the entire graph using a single matrix  $W \in R^{V \times V}$ , where each entry  $w_{ij}$  represents the strength of the connection between nodes  $i$  and  $j$ . The network topology is represented either by a non-weighted or weighted

adjacency matrix  $A$  that are respectively the same binarized matrix and the same matrix with zeros in place of values less, or less in modulus, than some certain threshold. Formally in case of non-weighted matrix,  $A \in R_+^{V \times V}$ , where each entry  $a_{ij}$  of the adjacency matrix is equal to 1 if an edge linking node  $i$  to node  $j$  exists, zero otherwise. In weighted adjacency matrices, the value of the entries corresponding to an existing link between nodes is  $a_{ij} = w_{ij}$ , zero otherwise.

Two important measurable graphs' properties, that have often a physiological meaning, are segregation and integration:

- Segregation: the more a network can be divided into subnets, the more segregated it is.
- Integration: the more closely interconnected the nodes in the graph are, the more integrated the network is.

Computing quantifiable measurements on graphs is important to evaluate objectively same aspects and to extrapolate information hidden in the adjacency matrix.

In fact, it is not just the graphical representation that makes graphs effective for expressing networks. One of the advantages of using a graph is related to the adjacency matrix, which enables the computation of quantitative and objective metrics known as graph's indices that can explain, detect and quantify aspects of the network that are challenging or impossible to identify by a visual examination of the graph. In fact, in the thesis work, I made use of graph's indices, to carry out some

statistical test, in order to evaluate one of the novel methods. In particular, the graph's indices used in my experiments are the following:

- Sum of the intra-group connections: given two groups of nodes, the sum of each connection's value belonging to that group.
- Sum of the inter-group connections: given two groups of nodes, the sum of each connection's value connecting nodes belonging to different groups.
- Weighted density of the intra-group connections: given two groups of nodes, the sum of each connection's value connecting nodes belonging to the same group divided by the sum of all the matrix's connections values.
- Weighted density of the inter-group connections: given two groups of nodes, the sum of each connection's value connecting nodes belonging to different groups divided by the sum of all the matrix's values.
- Divisibility: given two communities, it measures their segregation. Its value is within the range  $[0.5, 1]$ , where values close to 1 mean that the tested communities are completely unconnected. It is equal to the sum of the connections of the whole matrix divided by the sum between the connections of the whole matrix and the connections between nodes belonging to different communities. Lower divisibility results in greater higher integration, while higher divisibility equates to more segregation.
- Modularity: quantifies the propensity for particular subnetworks to merge into communities. A group of nodes is said to be in a

community when there are more connections between them than there are on average between any two nodes in the network. Positive modularity refers to the ability to divide a network into communities that function as a unit rather than as a collection of independent nodes. Instead, negative modularity describes how communities don't act like communities. Finally, when nodes are divided randomly, modularity takes values close to zero. Lower modularity promotes greater integration, while higher modularity promotes greater segregation.



# Chapter 2

## Traditional methods for estimating Wiener-Granger causality

Wiener-Granger causality, often simply referred to as Granger Causality (Seth, 2007), which has been used for more than 50 years, is now widely employed to analyse time series data in a variety of fields, from economics and finance to genomics and neuroscience (Shojaie & Fox, 2022).

Granger causality estimation relies on Multivariate Autoregressive Models (MVAR) (Kaminski & Blinowska, 2022).

An MVAR, for  $c$  channels, is expressed by the following formula:

$$X(n) = \sum_{k=1}^p A(k)X(n-k) + E(n)$$

where  $X(k)$  is a vector of signals,  $A(k)$  is a  $C \times C$  matrix of model coefficients,  $E(k)$ , in my implementation, is a  $c$ -size vector of white residual noises, and  $p$  is the model order (i.e.: the number of past samples taken into account in the regression).

Granger causality principle, in its bivariate form, states that for two time series, if the variance of the prediction error for the second time series  $Y$  is reduced by including past measurements from the first time series  $X$

in the linear regression model, then the first time series can be said to cause the second (Kaminski & Blinowska, 2022).

Formally:

$$\begin{cases} x[n] = \sum_{k=1}^p a_x[k]x[n-k] + e_x[n] \\ y[n] = \sum_{k=1}^p a_y[k]y[n-k] + e_y[n] \end{cases}$$

is the univariate Auto Regressive model, where  $a_x[k]$ ,  $a_y[k]$  are the model parameters for lag  $k$ ,  $p$  is the model order,  $e_x$  and  $e_y$  are the residuals associated with the model. The prediction error depends only on the past values of the own signal. Also,

$$\begin{cases} x[n] = \sum_{k=1}^p a_{xy}[k]x[n-k] + \sum_{k=1}^p b_{xy}[k]y[n-k] + e_{xy}[n] \\ y[n] = \sum_{k=1}^p a_{yx}[k]x[n-k] + \sum_{k=1}^p b_{yx}[k]y[n-k] + e_{yx}[n] \end{cases}$$

is the bivariate Auto Regressive model, where  $a_{xy}[k]$ ,  $a_{yx}[k]$  are the model parameters,  $p$  is the model order,  $e_{xy}$  and  $e_{yx}$  are the residuals associated with the model. In this case, the prediction error for each individual signal depends on the past values of both signals.

The prediction performances for both models can be computed by using the variances of the prediction errors. In fact, if  $var(e_{yx}) < var(e_x)$  then  $Y$  granger causes  $X$ .

Formally:

$$G_{y \rightarrow x} = \log_e \left[ \frac{\text{var}(e_x)}{\text{var}(e_{yx})} \right]$$

with  $G \approx 0$  if the past of  $Y$  does not improve the prediction of  $X$ , and  $G > 0$  otherwise.

The main perks of this approach are its statistical meaning and the directionality of the estimate. The main limitation lies on the fact that bivariate Granger causality does not consider information of other time series that may be causing both  $X$  and  $Y$ , resulting in spurious links (Faes, Vantieghem, & Van Hulle, 2022).

The Granger causality principle can be extended to an arbitrary number of channels by using as many MVAR models as the number of channels plus one ( $C + 1$ ).

In particular:

- $C$  “restricted” models  $A_1, A_2, \dots, A_C$  where in each model, the  $c$ -th equation, referred to the  $c$ -th channel is not included;
- 1 “full” model  $A_{full}$  where all the equation of all channels are present.

In this way, Granger causality takes the name of “Conditional” or “Multivariate” Granger Causality, and is computed in the following way:

$$G_{i \rightarrow j} = \log_e \left[ \frac{\text{var}(e_j^i)}{\text{var}(e_j^{full})} \right]$$

where  $var(e_j^i)$  is the  $j$ -th channel residuals' variance of the  $i$ -th restricted model, and  $var(e_j^{full})$  is the  $j$ -th channel residuals' variance of the full model.

In the case of Conditional Granger Causality, the formulation for the full model becomes:

$$\begin{cases} x_1[n] = \sum_{k=1}^p a_{11}[k]x_1[n-k] + \sum_{k=1}^p a_{12}[k]x_2[n-k] + \dots + \sum_{k=1}^p a_{1C}[k]x_C[n-k] + e_1[n] \\ x_2[n] = \sum_{k=1}^p a_{21}[k]x_1[n-k] + \sum_{k=1}^p a_{22}[k]x_2[n-k] + \dots + \sum_{k=1}^p a_{2C}[k]x_C[n-k] + e_2[n] \\ \vdots \\ x_C[n] = \sum_{k=1}^p a_{C1}[k]x_1[n-k] + \sum_{k=1}^p a_{C2}[k]x_2[n-k] + \dots + \sum_{k=1}^p a_{CC}[k]x_C[n-k] + e_C[n] \end{cases}$$

where  $A[n] = \begin{bmatrix} a_{11}[n] & \dots & a_{1C}[n] \\ \vdots & \ddots & \vdots \\ a_{C1}[n] & \dots & a_{CC}[n] \end{bmatrix} \in \mathbb{R}^{C \times C \times p}$  are the model parameters,

$i, j \in [1, 2, \dots, C]$ . (Cohen, 2014)

Another way to extend the Granger Causality principle to an arbitrary number of channels is by means of Directed Transfer Function (DTF) and Partial Directed Coherence (PDC), which are two connectivity measures based on the spectral version of the model (Kaminski & Blinowska, 2022).

Although all the methods mentioned in this Chapter depend on statistical assumptions that typically do not hold for EEG data, such as linearity and prior assumptions on connectivity being expressible as a relation between time series, these methods remain the best practice examples of

directed connectivity estimation. Even in case of Conditional Granger Causality, it is still plausible that some of the estimated graph edges are the result of a linear combination of unrelated causes. This is due to the fact that the signal obtained from a single electrode typically incorporates contributions from multiple sources (Faes, Vantieghem, & Van Hulle, 2022).

# Chapter 3

## Real and simulated encephalographic data

In order to estimate the performance of the proposed methods (see below) and to validate the physiological plausibility of the results, both synthetic and real data were used.

### 3.1 Synthetic EEG data

Real data does not provide any ground truth as cerebral connectivity can only be estimated and not measured directly; unless very invasive and hard-to-implement techniques are used, to measure intracranial potentials, which would provide a more precise measure of cerebral connectivity.

On the contrary, the use of synthetic data allows for ground truth and therefore the direct comparison of two or more methods.

In order to generate synthetic data, I made use of G-SEED (Anzolin, Toppi, Petti, Cincotti, & Astolfi, 2021), which enables the generation of pseudo-EEG data, with imposed connection patterns.

G-SEED toolbox generates data with linear dependences. Nowadays, it is still challenging to generate pseudo-EEG data with non-linear dependences, respecting the physiological behaviour of the brain. For this main reason and to avoid too much biased results towards my non-linear approach (that could not reflect real physiological assumption), I retained suitable the use of a linear pseudo-EEG generator. Another reason lies in the aim of the experiments, which is to have at least comparable results with traditional approaches, to assess feasibility of the first proposed method.

The toolbox allows to choose several parameters for data simulation. The main ones imposed for generating pseudo-EEG data for this work are reported below.

- Data length = 250
- Trials = 30
- SNR of additional noise =  $inf$
- Channels = 4
- Density = 40%
- AR model's values range =  $[-0.9, 0.9]$
- Model order = 6

I set the data length to 250 because this is a common number of samples per trial's window for most EEG experiments. A trial is a repetition of the experiment task.

Trials = 30 is enough to fit MVAR models and train the transformer-inspired network employed.

SNR of additional noise refers to a white Gaussian noise that can be added after pseudo-EEG generation (which is already affected by noise in generation phase). It has been set to *inf* for the experiment I will present in the following chapter. I tried also other values, which not let to significant difference in results.

The number of channels in real data is usually greater than 4 (e.g., 16, 32, 64, 128). However, 4 is enough to study the feasibility of the method, allowing to have shorter computational time and a better interpretability of results, both during experimentation phase and on presented results. The dimensionality of the data should not particularly affect the validity of the study. In order to apply the proposed method on higher-dimensional data, such as real data, only tuning of the model parameters to fit the new dimensions is necessary.

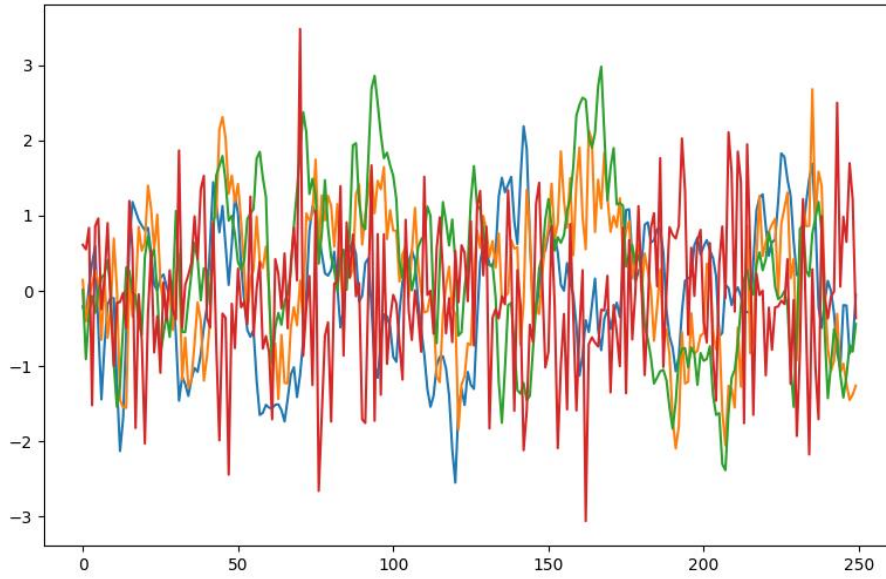
Density = 40% is a plausible value according to real data users' experience.

AR model's values range refers to the values of the autoregressive generation model. A range of  $[-0.9, 0.9]$  allows to have clear data in terms of strength of connections, while guaranteeing stability during generation.

I chose 6 as model order, as an arbitrary number that is quite small, allowing to have relatively short computational times.

The main strength of this toolbox lies in the fact that the simulated pseudo-EEG data show the same spectral and temporal properties as real EEG data.





*Figure 4: An example of the pseudo-EEG generated data.*

### 3.2 Real hyperscanning EEG data

Hyperscanning (Babiloni & Astolfi, 2014) is a technique that allows simultaneous analysis of the brain activity of two or more individuals by means of diagnostic techniques (e.g., EEG), employed while the individuals relate to each other in a social task.

The main reason why I chose hyperscanning data as benchmark for one of the methods I propose in this Thesis lies in the interpretability that this type of data provides. In fact, as the cerebral activity of the two or more subject taking part to hyperscanning experiment is simultaneously registered (by means of two or more EEG setups), it is possible to compute Granger causality extrapolating a unique matrix, carrying

information about the cerebral activity of all the subjects. In this way the matrix shows both intra and inter-subject causal interactions, leading to clusters that are clearly interpretable. In this way it is possible to infer physiological plausibility of a novel method for Granger causality estimation, simply looking at output matrices from a qualitative point of view. To the best of my knowledge, this is the first time that a transformer-based approach to causality estimate is applied on hyperscanning data.

### **3.2.1 Experiment description**

EEG data used in my work comes from an experimental study conducted by (Astolfi, et al., 2020). In their experiment, 32 male subjects (16 couples), between the ages of 18 and 30 (mean age = 25.28; SD = 4.39), participated in a computerized joint action task.

The Joint Action task, which was carried out through a computer game, was completed by each pair of individuals. The task involved controlling both sides (left and right) of a moving bar on which the virtual ball was placed in order to raise it from the bottom of the screen up to a target region at the top of the screen (the goal). If the proper balance was not maintained, the ball was free to roll down the bar (Figure 5). They added an obstacle to the centre of the screen to add complexity and applied a modified version of the paradigm proposed by Bosga and Meulenbroek (Bosga & Meulenbroek, 2007) by taking into account a solo condition and adding a non-human (PC) joint condition that is equivalent to the

human joint condition. As a result, there were three main conditions in the task.

In the Joint condition, the dyad worked on the same task together. Each participant used his right index finger to press a button to control one side of the virtual bar.

Similar to the Joint condition, in PC condition each subject used their right index finger to control one side of the virtual bar while the computer controlled the other side. This reflects joint coordination with a non-human agent.

In the solo experiment, each subject was required to complete the task on their own, utilizing their right index and middle fingers to control both sides of the virtual bar.

Additionally, the authors included a baseline condition in which the participants sat in front of a screen, watched a bar similar to the one used in the experiments move across the screen, and had to press buttons with the same fingers and timing as during the experiment, but with no relation to what was happening on the screen.

Both individuals played simultaneously while being recorded throughout all conditions, with no communication permitted other than what achieved by the joint task.

The subjects wore earplugs to prevent the noise of pushing buttons from facilitating their motor synchronization while they were seated face to face, separated from each other by a barrier that prevented them from looking at each other.

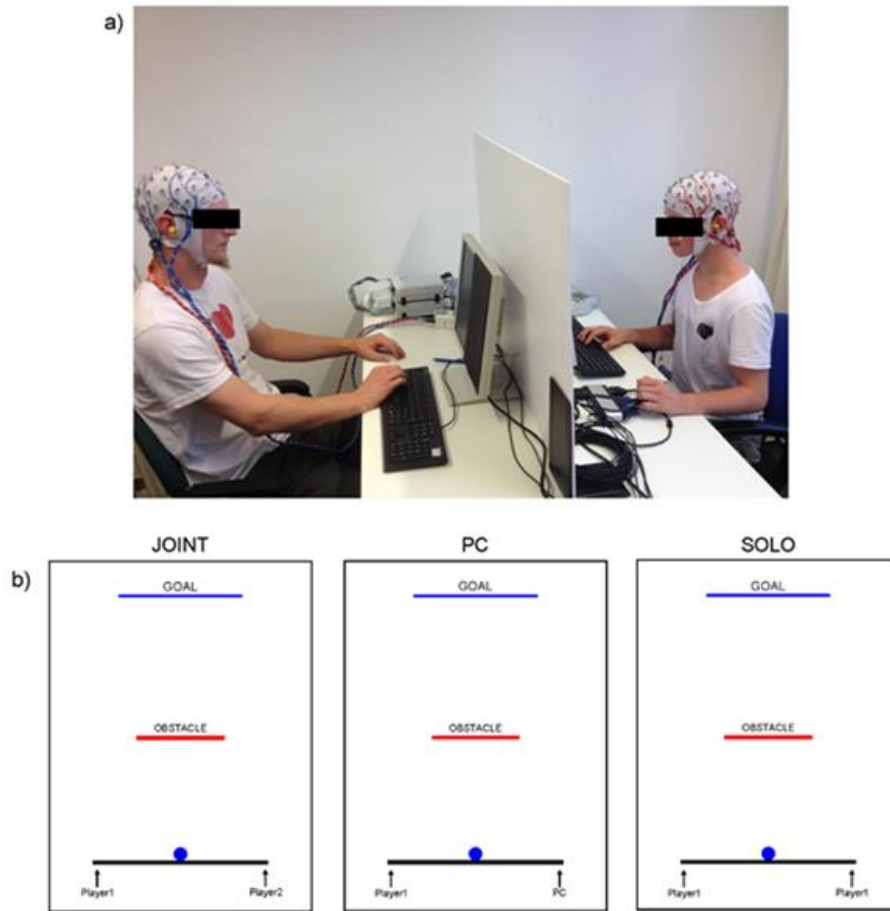


Figure 5:

a) Hyperscanning Experimental Setup.

b) Graphical representation of the three conditions included in the paradigm.

(From Astolfi et al, Raising the bar: Can dual scanning improve our understanding of joint action? *NeuroImage*, 216, 116813.2020)

### 3.2.2 Data processing and structure

Band-pass filters were applied to EEG recordings between 1 and 45 Hz. Ocular artifacts were eliminated using Independent Component Analysis (ICA). Only one ICA component (the one where the blink artifact was discovered) was eliminated from the estimated set. EEG traces were

divided according to the precise timing of the paradigm: in the Joint Condition, they kept traces when the two subjects shared control of the bar; in PC and Solo, they only took the time between the subjects' simultaneous start of the trial and the fastest subject's completion of the trial. These intervals were further divided into 1 second segments called epochs.

To remove the remaining artifacts, a semi-automatic approach based on an 80-volt threshold criterion was used. On average they removed less than 10% of the total trials per condition per subject. The number of epochs preserved for the three different experimental settings did not differ statistically from one another.

The data processing was completely carried out using Brain Vision Analyzer 1.0 (Analyzer, 2006).

# Chapter 4

## Proposed methods

In this chapter I will explain in detail the two novel methods and the traditional one used as reference.

### 4.1 Transformer models

As mentioned above, transform architectures (Vaswani, et al., 2017) were introduced in 2017 by a research group of Google Brain as a more efficient and performant alternative, in NLP tasks, to convolutional (Gu, et al., 2018) or recurrent neural networks (Yu, Si, Hu, & Zhang, 2019) in encoder-decoder configuration. Indeed, in the well-known article "Attention is all you need" (Vaswani, et al., 2017), the absence of convolutional and recurrence mechanisms is emphasised as a strength, enabling greater parallelization and less training time. Thanks to these characteristics as well as their performances with time series, transformer architectures are a promising choice for processing encephalographic data.

In this paragraph, I will briefly outline the architecture originally proposed in "Attention is all you need" (Vaswani, et al., 2017) for NLP tasks. All subsequent implementations of models based on the transformer architecture refer to this work.

### 4.1.1 Input and output

The transformer architecture in its original version has an encoder-decoder structure. The encoder aims to map a sequence of input symbols  $(x_1, \dots, x_n)$  into a sequence of continuous representations  $z = (z_1, \dots, z_n)$ . The decoder generates a sequence of output symbols  $(y_1, \dots, y_m)$  one element at a time. At each step the model takes the previously generated symbols as additional input for the generation of the next one, in an autoregressive manner.

The input and output sequences of symbols are converted respectively in and from  $d_{model}$  dimensional vectors using learned embeddings. The decoder output is converted to predicted next-token probabilities using learned linear transformation and softmax function.

### 4.1.2 Model architecture

Both the encoder and the decoder consist of stacked self-attention and fully connected layers. Figure 6 shows the encoder-decoder structure of the transformer.

In the encoder, each layer has two sub-layers. The first one is a multi-head self-attention mechanism, and the second is a position-wise fully connected feed-forward network. Around each of the two sub-layers, a residual connection is present, followed by layer normalisation.

The decoder follows the same structure with a major modification. Except for the first one, each layer has an additional sub-layer, which performs multi-head attention over the output of the encoder stack.

Another difference with the encoder lies in the first layer, where masked attention is used in the training phase to prevent predictions for a certain position from depending on future outputs. In practice, the method by which the decoder uses the encoder output depends on the various implementations of the transformer-inspired models.

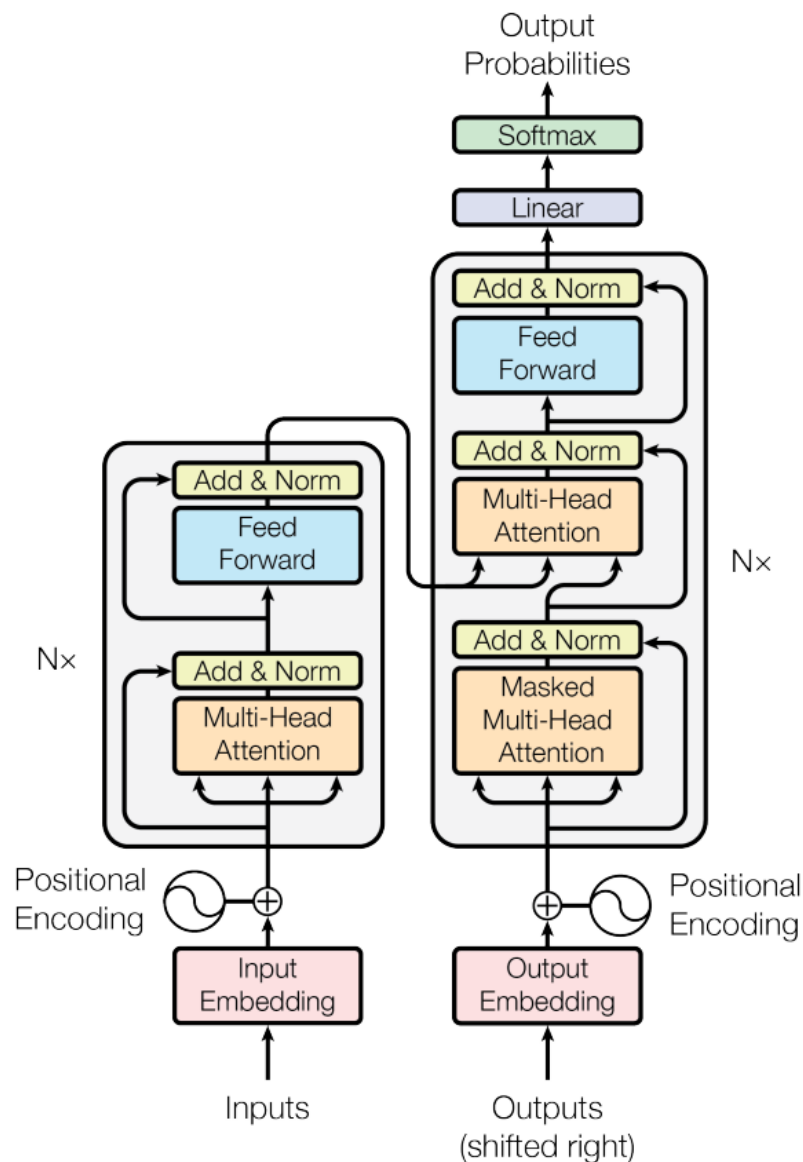


Figure 6: The transformer architecture as proposed in “Attention is all you need” (Vaswani, et al., 2017).



### 4.1.3 Multi-head attention mechanisms

An attention mechanism is a technique that enhances the values of some portion of the input data while diminishing others. The rationale is that it is often convenient to pay more attention to some small but important part of the data, depending on the context. Mathematically, an attention function maps a query vector and a set of key-value vector pairs to an output vector. It is calculated as a weighted sum of the values, with each value's weight determined by the query's compatibility function with its corresponding key.

In the case of scaled-dot product attention (Figure 7 (left)) compatibility function, as in (Vaswani, et al., 2017), the formula is the following:

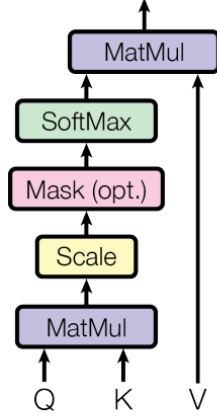
$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where  $Q$  and  $K$  are respectively the queries and the keys vector of dimension  $d_k$ ;  $V$  is the values vector.

The transformer architecture, instead of using a single attention mechanism per layer, uses more than one in parallel (Figure 7 (right)). It has proven useful to linearly project queries, keys, and values  $h$  times, using distinct linear projections learned of dimensions  $d_k$ ,  $d_k$ , and  $d_v$ , instead of executing a single attention function with keys, values, and queries of dimension. The attention function is then applied simultaneously to each of these projected versions of queries, keys, and values, producing  $d_v$  dimensional output values.

In this way, the model can compute attention to data from different sub-spaces of representation.

Scaled Dot-Product Attention



Multi-Head Attention

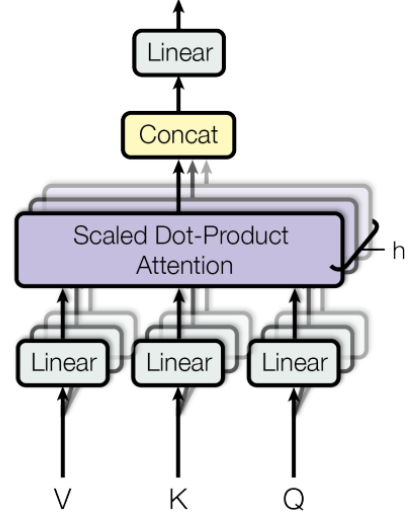


Figure 7: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel. From “Attention is all you need” (Vaswani, et al., 2017).

#### 4.1.4 Position-wise feed-forward networks

Each layer of our encoder and decoder has a fully connected feed-forward network  $FFN$  as the last sublayer, which is applied to each position independently and keeping the same learned weights. This network is composed of two linear transformations with a  $ReLU$  activation in between.

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Input and output have the same dimension  $d_{model}$  while the inner layer has dimensionality  $d_{ff}$ .

### 4.1.5 Positional Encoding

In order to provide the model with information regarding the relative or absolute position of the tokens, it is necessary to add this information to the input data, since our model has no recursive mechanisms, no convolution, and no way to extrapolate the order of the tokens in the sequence. For this reason, positional encodings are added to the input embeddings at the base of the encoding and decoding stacks. Positional encodings and embeddings can be added together since they share the same model size. They can be either learned from the model or imposed, but in the experiments carried out in this work (Vaswani, et al., 2017), no significant performance difference was found between the two paradigms.

The following chapter discusses the transformer-inspired architecture employed in the method proposed by this thesis, which, in contrast to the one just presented, is suitable for dealing with time series, particularly multivariate ones.

## 4.2 The Spacetimeformer architecture

As already noted, building on its success in NLP, new architectures strongly inspired by transformers have seen light in recent years, making it possible to apply the same concept in fields as computer vision and time series forecasting.

The Spacetimeformer (Grigsby, Wang, & Qi, 2021) gets its name from its peculiarity of being able to directly pose attention to data both spatially (i.e., channel axis) and temporally. In this way, the flow of information within the network follows patterns that are less convoluted and that make the model more easily fittable by optimizers.

The Spacetimeformer is used in both methods I propose as a prediction model for the next sample. However, in the two methods I use the Spacetimeformer in very different ways:

1. In the first case I use it instead of an MVAR in the calculation of Conditioned Granger causality,
2. In the second, I exploit the explicability of the values of attention computed at the prediction stage in order to construct the causality matrix. This is possible because of the Spacetimeformer's peculiarity of computing attention on the channel axis in a direct manner, making the computation of attention values more explainable, so that they can be used by the second proposed method to compute the causality matrices.

For the reasons listed above, the Spacetimeformer is both suitable for dealing with multivariate EEG data and adequate for the needs of the two proposed methods.

In this paragraph I explain the most relevant novelties introduced by Grigsby et al. (Grigsby, Wang, & Qi, 2021) in the Spacetimeformer, which make this architecture so peculiar and suitable for my purposes.

To a first approximation, the Spacetimeformer converts the multivariate sequence of dimension  $(channels \times time)$  into a spatiotemporal

sequence of dimension  $(channels \cdot time)$ , where each input token reflects the value of a single variable at a specific time. Then, over this extended sequence, a transformer model can jointly understand how space, time, and value information interact, computing attention on the tokens of this extended sequence.

### 4.2.1 Input sequence

As explained by Grigsby et al. (Grigsby, Wang, & Qi, 2021), starting from a  $d$ -dimensional embeddings of the sequence:

$$(x_{T-c}, y_{T-c}), \dots, (x_T, y_T), (x_{T+1}, 0_{T+1}), \dots, (x_{T+h}, 0_{T+h}),$$

it is obtained the result expressed in matrix form as  $Z \in \mathbb{R}^{(c+h) \times d}$ .<sup>41</sup> Flattening each multivariate  $y_t$  vector into  $N$  scalars associated to its timestamp  $x_t$ , we get a new sequence:

$$((x_{T-c}, y_{T-c}^0), \dots, (x_{T-c}, y_{T-c}^N), \dots, (x_T, y_T^0), \dots, (x_T, y_T^N), \\ (x_{T+1}, 0_{T+1}^0), \dots, (x_{T+1}, 0_{T+1}^N), \dots, (x_{T+h}, 0_{T+h}^N)).$$

Embedding this new longer sequence results in a  $Z' \in \mathbb{R}^{N(c+h) \times d}$ . Feeding  $Z'$  through a Transformer the attention matrix  $A(Z', Z') \in \mathbb{R}^{N(c+h) \times N(c+h)}$  represents a spatiotemporal graph with direct paths between each variable at each time sample.

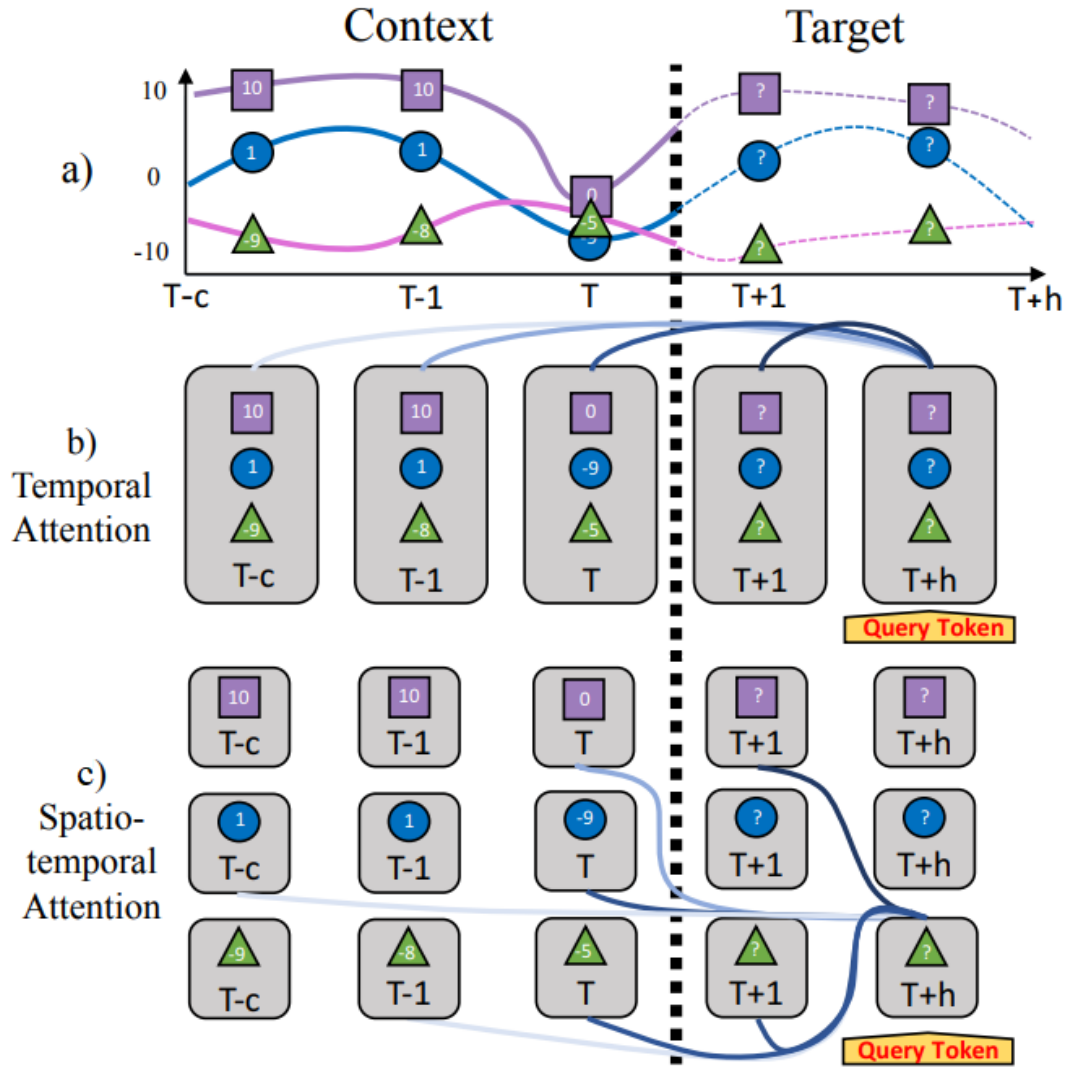


Figure 8: Types of attention between tokens in multivariate forecasting. (a) A three-variable sequence of three context points and two target points to be forecasted. (b) Temporal attention, where each token contains all the three variables. Increased attention. (Grigsby, Wang, & Qi, 2021)

#### 4.2.2 Spatiotemporal embedding

The transformer must be given the freedom to correctly interpret the variable each token derives from, as well as its time and scalar value, after the input variables have been flattened. Transformers are

permutation invariant, which means they are unable to automatically interpret the order of input tokens. A position embedding, typically with a fixed sinusoidal pattern, is added to the tokens to address this. Using a Time2Vec layer, positional embeddings are learned passing a representation of absolute time through sinusoidal patterns of learned offsets and wavelengths. This makes interpreting periodic patterns easier. However, in my case, there is no need to model periodic patterns, especially over long periods of time, thus a relative order index is added to clearly distinguish temporal samples in a small time-window.

In order to distinguish also the difference between the variables at each of these times, a “variable embedding” is added to each token. The variable embedding refers the time series (i.e., channel) of which each token is a part. This mechanism is implemented as a standard embedding layer that maps the integer index of each series to a higher-dimensional representation by causing the variable representations to be randomly initialized and learned end-to-end during training.

After having processed and embedded the input sequence by also exploiting spatiotemporal embedding, the transformer model (which I already treated in the previous section) processes it, and outputs a prediction. Since flattening the input sequence may results in a very long one, the model can use additional mechanisms to treat with long sequences, such as down-sampling convolutions and local attention mechanisms (Niu, Zhong, & Yu, 2021), depending on the configuration chosen, based on necessities.

With the words “cross-attention” I will refer hereafter to the attention computed between the encoder output and each decoder layer, as already explained in the previous section.

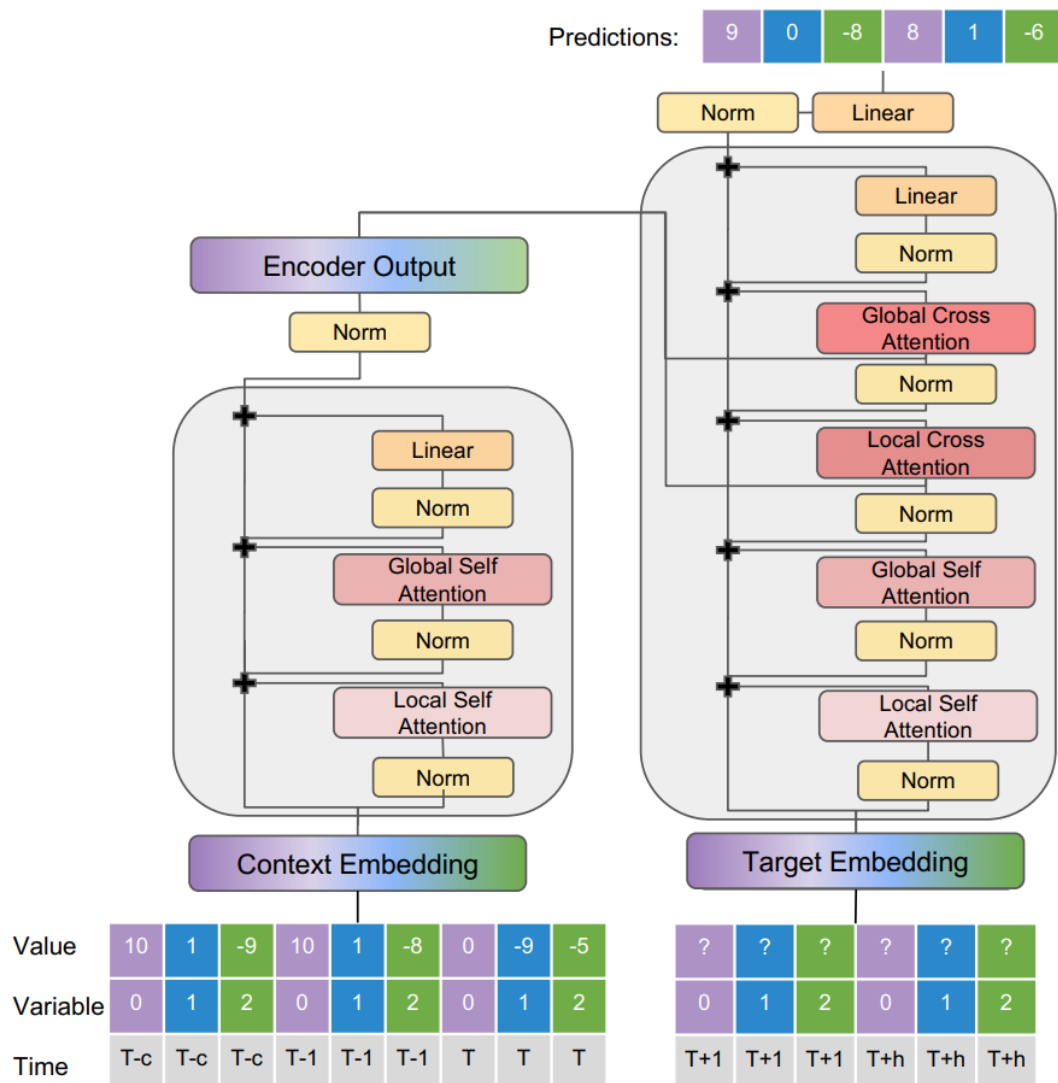


Figure 9: The Spacetimeformer architecture applied to the sequence shown in Figure 8a. (Grigsby, Wang, & Qi, 2021)



## 4.3 Experiments

In this section I explain in detail the two new methods I propose in this thesis, as well as the method used as a reference for the first proposed method.

After the generation of synthetic EEG data, the work has been divided in three main steps:

1. Estimation of conditional Granger causality using an MVAR as a predictive model, on synthetic EEG data, to be used as a traditional linear reference method for comparison with the first proposed method, to assess its plausibility.
2. Estimation of conditional Granger causality using the Spacetimeformer as a predictive model.
3. Comparison between the matrices estimated with the two methods with the ground truth in common.
4. Carry out a statistical test on the collected data.

For the second proposed method the main steps are:

1. Estimation of causality matrices computed using the attention values calculated by the Spacetimeformer, on the two conditions "Solo" and "Joint" separately.
2. Graph's indices computation on the matrices in output of the first step.
3. Statistical tests to assess method physiological validity.

### 4.3.1 Reference method for CGC estimation applied to synthetic data

As already anticipated, in order to assess the plausibility of the first proposed method I chose to generate synthetic data using an imposed autoregressive model. In this way, in order to carry out a statistical test to assess plausibility of the first proposed method, I was able to compare the causality matrices generated by the traditional reference method (i.e., Conditional Granger causality) and the ones obtained by the first proposed method with a common ground. I will treat about the first novel method in the following section.

I chose CGC as reference method mainly for two reasons:

1. The first proposed method works in the temporal domain, and to minimise the possible bias towards one of the two methods I chose the state-of-the-art method amongst the non-spectral ones.
2. Being based on MVAR linear models, is suitable for dealing with data having linear causal dependencies. This is a point against the proposed non-linear method, although the purpose of the comparison is to ensure the reasonableness of the method, thus having comparable performance.

In the rest of this sub-section I report the operative steps of the reference method.

For every repetition I used  $O_m = 2$  different optimization methods (i.e., Yule-Walker and Vieira-Morf (Marple Jr & Carey, 1989)) repeating both the experiments  $R = 10$  times. I imposed as MVAR's order dimension

(i.e., 6) the same of the ground truth AR model  $GT$  used during pseudo-EEG generation. The choice of imposing the model order instead of estimating it from the data has been made for reducing the variability of the results, with more comparable results, with less repetitions. I obviously made the same choice for the proposed method as well, for the same reason.

For each repetition  $r$  of the experiment and for each optimization method  $o_p$ , I fitted an MVAR model  $A_{full}$  and as many MVAR models  $A_{1,2,\dots,C}$  as the number of channels  $C = 4$  defined as in Chapter 3.

For each MVAR model  $A_i$  and for each channel  $c_j$  I computed the residual  $e_j^i$ . Applying the Granger formula for each pair of channels  $(c_i, c_j)$  I obtained the causality matrix  $G \in \mathbb{R}^{C \times C}$ .

$$G_{i \rightarrow j} = \log_e \left[ \frac{\text{var}(e_j^i)}{\text{var}(e_j^{full})} \right]$$

Before comparing the causality matrix obtained with the ground truth, I processed it, for three reasons:

1. To get rid of some “noise” of the estimation process (i.e., negative numbers in the Granger causality matrix are nonsignificant)
2. To make it possible to apply the Jaccard Distance function (Ioffe, 2010), which works only with positive matrices.

I processed the two matrices as follows:

1. Negative numbers were put to zero after matrix standardization.

2. Normalization between 0 and 1 after the negative numbers' thresholding.

Once obtaining the processed causality matrix  $G$ , I compared it with the ground truth (i.e., the AR model used for generating pseudo-EEG data). As a metric, I used the Jaccard Distance (Ioffe, 2010).

$$Jaccard\ Distance\ (G, GT) = 1 - \frac{\sum_i \sum_j \min(g_{ij}, GT_{ij})}{\sum_i \sum_j \max(g_{ij}, GT_{ij})}$$

The Jaccard Distance provides a topology measure of how different the two matrices are.

Experimental results are reported and discussed in the following Chapter.

### **4.3.2 Novel method for CGC estimation applied to synthetic data**

Assuming that, as already mentioned, spatio-temporal causal dependencies have a nonlinear nature, it is interesting to model them with a nonlinear method that is designed to deal with data having such characteristics (ie., spatio-temporal causal relationships, such as the Spacetimeformer).

The basic idea of the first proposed method is to replace the MVAR linear model with a Spacetimeformer model. The assumption behind this choice is that using a non-linear forecasting model, computed residuals must reflect the non-linearities of the model, implying in output of the

pipeline (after the Granger formula application), a more precise Causality matrix, that captures the non-linearities present in the data.

Data is divided in two subsets:

1. Training set, with 70% of trials, is used to fit the Spacetimeformer model.
2. Validation set, with 30% of trials, is used as metric to choose the best model's weights to consider.

The residuals are then computed on the union of the two sets.

Data is not pre-processed in any way, it is standardized along trials and channels axes and subsequently divided into as many windows as *number (#) of windows*:

$$\# \text{ windows} = (\# \text{ sample} \cdot \# \text{ trials}) - \# \text{ context points}.$$

Each window is long  $\# \text{ context points}$  (6) +  $\# \text{ target points}$  (1) samples.

For the rest of the thesis, I will refer to the words "number of" using the symbol "#".

It is possible to choose several hyper-parameters and configurations in the Spacetimeformer architecture. Every choice made was based on my own reasoning, and on experiments conducted both by me and those reported by the authors Grigsby et al. (Grigsby, Wang, & Qi, 2021). The most relevant ones are reported and explained below:

- $d_{model} = 15$ ,  $d_{ff} = 40$ ,  $d_k = 4$ ,  $d_v = 4$ ,  $\# \text{ heads} = 3$ ,  
 $\# \text{ encoder layers} = 2$ ,  $\# \text{ decoder layers} = 2$ :

This choice of hyperparameters, whose meaning has already been explained in the previous sections, is essentially a compromise between performances and computational time. As basis they follow approximately the same proportions between each other as the average set of hyperparameters used in experiments conducted by Grigsby et al. (Grigsby, Wang, & Qi, 2021).

- *# context points = 6, # target points = 1:*

As for the reference method, the optimal order of the model was not estimated, but rather imposed a priori equal to the value of ground truth's order (e.g., *# context points = 6*). The number of *target points* is set to 1, as for the MVAR traditional approach.

- *batch size = 64, # warm-up steps = 1000,*  
*base learning rate =  $5 \cdot 10^{-4}$ , initial learning rate =  $5 \cdot 10^{-7}$*   
*learning rate decay = 0.5, epochs = 50,*  
*optimizer = AdamW, loss function = MSE:*

Learning rate is not constant but varies over time (Figure 10) according to the learning rate annealing method (Nakamura, Derbel, Won, & Hong, 2021), which contributes to more stable and faster learning. The method consists of starting with a very small *initial learning rate*, incrementing it until the *base learning rate* is reached. Then, the learning rate is decremented using a scheduler or, as in my case, decreased it by a factor of *learnig rate decay* as the validation loss starts to plateau. “*# warm-up steps*” refers to the amount of training steps needed for reaching the *base learning rate*.

The rationale behind the warm-up phase is based on two observations:

1. As the model parameters are randomly initialized, the initial model is very far from the ideal solution. Thus, an excessively large learning rate may cause numerical instabilities.
2. Carefully training a model in the first few epochs allows to apply a greater learning rate during the middle of the training, leading to a better regularization.

I chose this combination of parameters because are empirically the ones which ensure to fit the model in  $epochs = 50$  at a reasonable level of fitting (considering computational times to fit as many models as the number of channels, plus one). See Figure 10.

As optimizer and loss function, I chose AdamW and Mean Squared Error (MSE) respectively, as suggest by Grigsby et al. (Grigsby, Wang, & Qi, 2021).

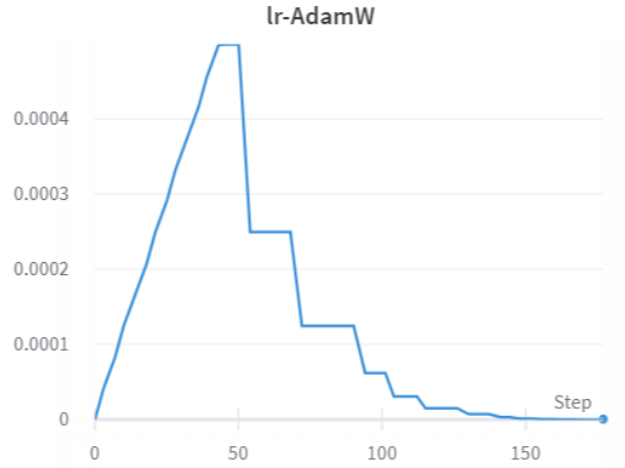


Figure 10: Learning rate during training one of the Spacetimeformer models of the first proposed method. Note the behaviour of the annealing method.

All random operations, such as initializing the network weights and shuffling the data are performed with a constant seed = 0. These two measures are important, as the different models need to be as comparable as possible in order to make an accurate causality estimate.

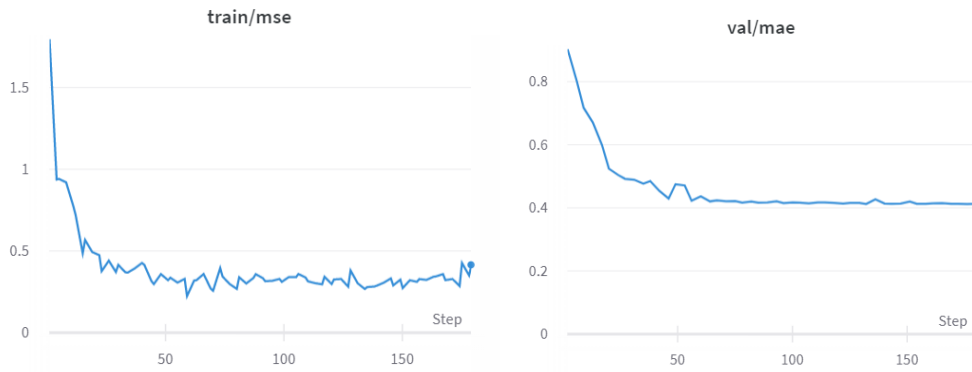


Figure 11: Training (left) and validation (right) losses (MSE) during training of one of the Spacetimeformer models of the first proposed method.



After having trained all the models needed, the method pipeline finishes computing the causality matrix  $G$  in the same way as explained for the reference traditional method.

Even in this case, each output matrix  $G$ , after processing (as for the reference method), is compared with the associated ground truth  $GT$  by means of the Jaccard Distance.

Experimental results are reported and discussed in the following Chapter.

### **4.3.3 Novel attention-based method for causality estimation applied to hyperscanning data**

The core of the novelty of this method lies in two main aspects:

1. Using attention computed from a model based on transformer architecture as a causality estimate of EEG data, in this case from the Spacetimeformer.
2. Applying a deep learning model to hyperscanning data.

My hypothesis is that the attention computed by the model may represent some form of causality. More extensively, the assumption is that, the model, in order to minimise the loss on a forecasting task (i.e., to forecast future samples having as input a window of multivariate past samples), is incentivised to learn to pay more attention to the parts of the input that are more significant for the correct estimation of the future values of each variable; in terms of both lags and channels, indirectly respecting the concept of Granger causality.

As for the first proposed method, data is divided in two subsets (i.e., training set and validation set) with the same proportions (i.e., 70% and 30%).

The attention is then computed on the union of the two sets.

In this case data is sub-sampled by a factor of 2 (i.e., from 250 Hz to 125 Hz) for physiological assumptions. Additionally, data has been derived to ensure stationarity and standardized along trials and channels axes and subsequently divided into windows of  $\# \text{context points}$  (10) =  $+ \# \text{target points}$  (1) samples.

It is possible to choose several hyper-parameters and configurations in the Spacetimeformer architecture. Every choice made was based on my own reasoning, and on experiments conducted both by me and those reported by the authors Grigsby et al. (Grigsby, Wang, & Qi, 2021). The most relevant ones are reported and explained below.

As for the previous method, I will discuss the choice made in terms of hyper-parameters of the Spacetimeformer. Only those that are not in common with the first proposed method are reported, exception made for the number of the encoder and decoder layers:

- $\# \text{encoder layers} = 2, \# \text{decoder layers} = 2$ :

The number of layers of encoder and decoder is limited mainly to force the model to concentrate knowledge on two layers only. The rationale is that, for my assumptions, more layers are available and more difficult and convoluted is to aggregate the attention computed by each layer. Conversely too few layer lead to

underfitting, failing to capture some non-linearities that are assumed to have real data, especially hyperscanning EEG data.

- *# context points = 10:*

The model order was chosen according to the physiological assumptions.

- *learning rate decay = 0.7, epochs = 90:*

In this case I chose to train the model at a finer level, for making sure to get the most accurate results.

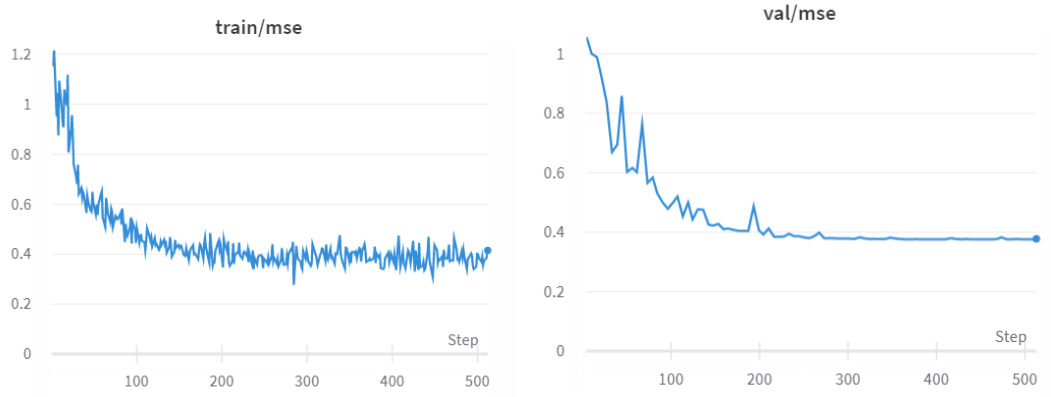


Figure 12: Training (left) and validation (right) losses (MSE) during training of one of the Spacetimeformer models of the second proposed method.

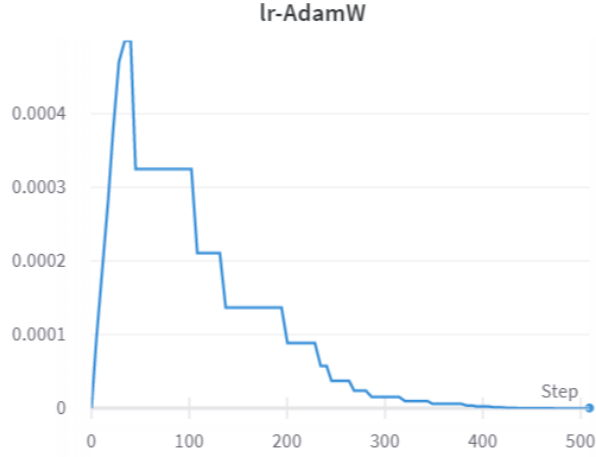


Figure 13: Learning rate during training one of the Spacetimeformer models of the second proposed method. Note the behaviour of the annealing method.

Once the model has been trained, the pipeline finishes with the computation of the causality matrix.

The first step is to average the computed attention matrices along the heads' dimension, obtaining  $A_s \in \mathbb{R}^{(C \cdot S) \times (C \cdot S)}$  for self-attention layers and  $A_c \in \mathbb{R}^{C \times (C \cdot S)}$  coming from a cross-attention layer, where  $C$  is the number of channels and  $S$  is the model order (i.e., the number of samples for each window, the temporal dimension). Averaging along the heads means to aggregate information computed on different projections of the same input (the role of the multiple heads) with a linear operation. I retained this an acceptable approximation to aggregate data. However, experimenting with different aggregation functions can certainly be the object of future extensions of this work.

After averaging data across the  $S$  dimensions for both cases, I end up with  $A_s \in \mathbb{R}^{(C \times C)}$  and  $A_c \in \mathbb{R}^{(C \times C)}$ . In this case I retain the average a

proper aggregation function as the purpose is to get rid of time attention, as we are interested in causal dependences between channels.

In my experiments, I tried different methods of aggregating data coming from the computed attention matrices over all the attention layers. Some of them are actually not aggregations, but directly the matrices themselves.

1. Cross-attention on Layer 0.
2. Cross-attention on Layer 1.
3. Average of cross-attentions.
4. Element-wise product between cross-attentions.
5. Self-attention on Layer 0.
6. Self-attention on Layer 1.
7. Average of self-attentions.
8. Element-wise product between self-attentions.
9. Average of all the attention on every layer.
10. Element-wise product between all the attention on every layer.
11. Element-wise product between the averages of cross and self-attention.

The method was applied to all the dyads of subjects present in the dataset of hyperscanning EEG data already described.

Each couple of attention matrices  $A_s, A_c$  was processed according with the following algorithm:

1.  $A_s^s, A_c^s = \text{standarization}(A_s), \text{standarization}(A_c);$

2.  $minimum = \min(A_s^s, A_c^s)$  (the minimum scalar value of the two matrices);
3.  $A_s^p, A_c^p = (A_s^s + minimum), (A_c^s + minimum)$ .

where  $A_s^p, A_c^p$  are the processed resulting matrices.

In this way, the distributions of the obtained matrix values are centred at the same point (i.e., they have the same mean), but devoid of negative numbers, which would lead to problems in computing the indices. For example, the densities would not be computable because the denominator, which is the sum of the matrix values, would be 0. Having the same mean and unitary standard deviation, the resulting matrices are more comparable.

After that, the following graph's indices (as introduced in Chapter 1) were computed:

1. Sum of the intra-subject values.
2. Sum of the inter-subject values.
3. Weighted density of the intra-subject connections.
4. Weighted density of the inter-subject connections.
5. Divisibility.
6. Modularity.

In order to evaluate the physiological plausibility of the method, I performed a statistical test on the graph's indices computed, which is presented in the following Chapter.

# Chapter 5

## Results

I carried out a statistical analysis for both the proposed methods, to assess their accuracy and physiological plausibility, respectively.

### 5.1 Statistical test on the novel method for CGC estimation applied to synthetic data

In order to compare the first proposed method with the reference one, I carried out four paired one-side t-tests (Hsu & Lachenbruch, 2014) on the resulting data from Conditioned Granger causality computation with both the reference approach and the proposed one.

In particular, for the reference approach, I tested both Yule-Walker and Vieira-Morf MVAR optimization methods.

In addition, the four tests are repeated for both types of matrix processing:

1. Negative numbers to zero thresholding after matrix standardization.

2. Normalization between 0 and 1 after Negative numbers to zero thresholding.

Figure 14 depicts a comparison between some Granger causality matrices computed with each method. In the case showed, none of the methods was able to detect the weakest connection. The other three connections were properly detected by Yule-Walker GC and Spacetimeformer GC. Vieira-Morf GC was able to detect the topology of these three connections but not their values.

For every repetition of the same experiment, with a different ground truth  $GT$ , the Jaccard Distance between the matrix obtained with each method and  $GT$  was computed. The results collected are reported in Table 1.



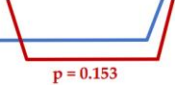
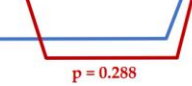
On these data I performed a t-test to investigate the presence of significant differences between the new method and both implementations of the reference one. In all cases, the test did not reveal any statistical significance. This result is in line with the expectations and goals for this method. Applying a non-linear method to linear data, it is reasonable to expect comparable, if not worse, performance than a linear method. However, the fact that there is no statistical evidence that either method performs better than the other, it can be concluded that the proposed method may be a good basis for future studies.

Regarding the choice to standardize or normalize between 0 and 1 the output matrices, from a qualitative analysis of the results, there seems to be no significant difference in performance. It is therefore slightly



preferable to normalize between 0 and 1, because thresholding to zero before any other manipulation is more conservative, since it affects only to the values that are surely the result of estimation noise, being negative.

Table 1: Each row represents a repetition of the same experiment with a different ground truth. Each column represents a different experiment. Each value is the Jaccard Distance between the method of the associated column and the ground truth of data associated to the row. Statistical tests returned no significant difference between the proposed methods and the two baselines, as is highlighted by blue and red links. The red and green values are the maximum and minimum average values, respectively, for each type of normalization.

	Negative to zero thresholding after standardization			Negative to zero thresholding before normalization		
	Yule-Walker	Vieira-Morf	Spacetimeformer	Yule-Walker	Vieira-Morf	Spacetimeformer
	0.35	0.35	0.35	0.29	0.30	0.32
	0.59	0.57	0.51	0.49	0.48	0.48
	0.25	0.26	0.23	0.23	0.24	0.23
	0.42	0.30	0.37	0.39	0.30	0.37
	0.08	0.22	0.14	0.10	0.24	0.12
	0.59	0.57	0.56	0.49	0.48	0.51
	0.25	0.26	0.26	0.23	0.24	0.27
	0.42	0.30	0.38	0.39	0.30	0.39
	0.08	0.22	0.13	0.10	0.24	0.10
	0.22	0.24	0.14	0.19	0.29	0.20
MEAN	0.323	0.330	0.308	0.291	0.312	0.298
STD. DEV	0.18	0.13	0.15	0.15	0.09	0.14
						
						

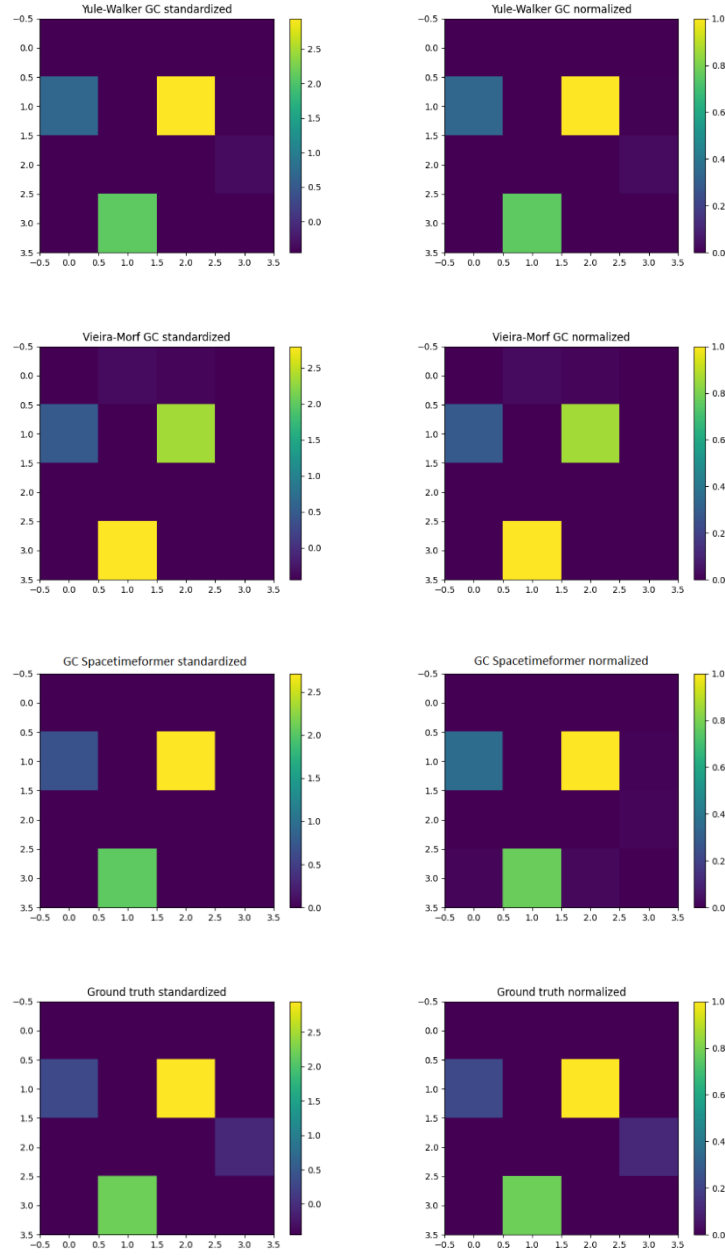


Figure 14: Set of example plots. Starting from above, the first column refers to standardized and then zero-thresholded matrices, preserving positive numbers. The second one refers to matrices that are firstly zero-thresholded preserving positive numbers and then normalized between 0 and 1. In the case presented, none of the methods was able to detect the weakest connection (on the right). The other three connections were properly detected by Yule-Walker GC and Spacetimeformer GC. Vieira-Morf GC was able to detect the topology of these three connections but not their values.

## 5.2 Statistical test on the novel attention-based method for causality estimation applied to hyperscanning data

In order to evaluate the quality of the second novel method, I carried out a statistical test in addition to a qualitative analysis of the results (that are reported in Table 3). For each type of attention matrix (or aggregation of them) and for each graph's index, I performed a paired one-sided t-test to statistically compare the two conditions (i.e., "Solo" and "Joint").

The results of all the t-test carried out are reported in Table 2.

The box plots referred to the significance found are reported in Figure 15, Figure 17, Figure 19. Associated to each box plot, there is a figure showing the related average (cross-subjects) output causality matrices for the two conditions (i.e., "Solo" and "Joint"), on which the box plots are based; (Figure 16, Figure 18, Figure 20).

The numerous significances found are physiological and in line with both theoretical assumptions and previous works on this dataset (Astolfi, et al., 2020). The multi-subject matrices clearly showed two clusters (corresponding to the nodes belonging to the two subjects), and this division is stronger in the individual condition (Solo) than in the social

condition (Joint), with a clear physiological interpretation in terms of inter-subject connectivity.

To quantify this aspect, graph indices were used. From the results of the statistical analysis, a higher integration can be observed in the Joint condition than in the Solo condition, as testified by the higher inter-subject density and weight (measures of integration) and by the lower Divisibility and intra-subject density (measures of segregation). These results confirm the quality of the novel proposed method, based on Spacetime-former attention computation.

Table 2: Results obtained from paired one-sided  $t$ -tests computed for each type of attention matrix or aggregation of them. Each pair of columns ( $t$ -stat.  $p$ -value) is referred to a graph's index. As for the aspects related to the choice of attention matrices to be used as a new connectivity estimator, it is interesting to note that statistical significance was found only for the aggregations of matrices or attention matrices from the encoder layers (i.e., self-attention layers).

Matrix type	Graph's index		Sum intra-sub.		Sum inter-sub.		Density intra-sub.		Density inter-sub.		Divisibility		Modularity	
	t-stat.	p-value	t-stat.	p-value	t-stat.	p-value	t-stat.	p-value	t-stat.	p-value	t-stat.	p-value	t-stat.	p-value
1. Cross-attention on Layer 0	1.69	0.1119	-1.69	0.1119	1.47	0.1613	-1.47	0.1613	1.47	0.1613	0.34	0.7381		
2. Cross-attention on Layer 1	-0.99	0.3366	0.99	0.3366	-0.90	0.3828	0.90	0.3828	-0.85	0.4089	-0.91	0.3752		
3. Average of cross-attentions	1.14	0.2704	-1.14	0.2704	1.24	0.2340	-1.24	0.2340	1.27	0.2251	0.22	0.8262		
4. Element-wise product between cross-attentions	-0.27	0.7902	0.27	0.7902	-0.35	0.7348	0.35	0.7348	-0.30	0.7705	-0.42	0.6771		
5. Self-attention on Layer 0	1.05	0.3121	-1.05	0.3121	1.28	0.2191	-1.28	0.2191	1.30	0.2143	-0.47	0.6433		
6. Self-attention on Layer 1	2.40	<b>0.0296</b>	-2.40	<b>0.0296</b>	2.59	<b>0.0204</b>	-2.59	<b>0.0204</b>	2.62	<b>0.0192</b>	1.64	0.1224		
7. Average of self-attentions	2.04	0.0592	-2.04	0.0592	2.28	<b>0.0377</b>	-2.28	<b>0.0377</b>	2.31	<b>0.0355</b>	1.08	0.2953		
8. Element-wise product between self-attentions	2.29	<b>0.0369</b>	-2.29	<b>0.0369</b>	2.61	<b>0.0197</b>	-2.61	<b>0.0197</b>	2.66	<b>0.0180</b>	1.28	0.2209		
9. Average of all the attention on every layer	1.78	0.0950	-1.78	0.0950	1.82	0.0894	-1.82	0.0894	1.84	0.0854	0.49	0.6298		
10. Element-wise product between all the attention on every layer	0.34	0.7364	-0.34	0.7364	0.15	0.8810	-0.15	0.8810	0.34	0.7356	-0.69	0.5007		
11. Element-wise product between the averages of cross and self-attention	1.93	0.0725	-1.93	0.0725	1.91	0.0757	-1.91	0.0757	1.95	0.0702	0.14	0.8928		

Table 3: Results of graph’s indices application to the causality matrices computed with the novel attention method. Each row is related to a different type of attention matrix or aggregation of them. Each column is related to the mean or standard deviation of a different graph’s index, applied to each condition (i.e., “Solo” and “Joint”).

Matrix type	Sum intra-subject				Sum inter-subject			
	MEAN		STD		MEAN		STD	
	Solo	Joint	Solo	Joint	Solo	Joint	Solo	Joint
1. Cross-attention on Layer 0	623	603	123	120	510	530	132	130
2. Cross-attention on Layer 1	447	461	41	50	262	248	94	88
3. Average of cross-attentions	775	758	114	125	594	611	123	109
4. Element-wise product between cross-attentions	323	326	44	34	122	120	38	36
5. Self-attention on Layer 0	1103	1096	202	210	1038	1045	208	198
6. Self-attention on Layer 1	905	872	114	120	762	795	133	105
7. Average of self-attentions	1137	1109	138	154	993	1021	170	140
8. Element-wise product between self-attentions	720	688	91	105	580	613	125	89
9. Average of all the attention on every layer	1085	1060	128	119	871	897	123	124
10. Element-wise product between all the attention on every layer	273	270	49	34	90	93	39	34
11. Element-wise product between the averages of cross and self-attention	694	667	138	139	477	505	132	124
Matrix type	Density intra-subject				Density inter-subject			
	MEAN		STD		MEAN		STD	
	Solo	Joint	Solo	Joint	Solo	Joint	Solo	Joint
1. Cross-attention on Layer 0	0.553	0.535	0.036	0.029	0.447	0.465	0.036	0.029
2. Cross-attention on Layer 1	0.641	0.660	0.084	0.082	0.359	0.340	0.084	0.082
3. Average of cross-attentions	0.568	0.555	0.030	0.026	0.432	0.445	0.030	0.026
4. Element-wise product between cross-attentions	0.727	0.734	0.075	0.063	0.273	0.266	0.075	0.063
5. Self-attention on Layer 0	0.516	0.512	0.013	0.009	0.484	0.488	0.013	0.009
6. Self-attention on Layer 1	0.545	0.523	0.037	0.015	0.455	0.477	0.037	0.015
7. Average of self-attentions	0.535	0.521	0.026	0.010	0.465	0.479	0.026	0.010
8. Element-wise product between self-attentions	0.557	0.529	0.045	0.017	0.443	0.471	0.045	0.017
9. Average of all the attention on every layer	0.555	0.542	0.024	0.018	0.445	0.458	0.024	0.018
10. Element-wise product between all the attention on every layer	0.755	0.751	0.092	0.061	0.245	0.249	0.092	0.061
11. Element-wise product between the averages of cross and self-attention	0.596	0.571	0.040	0.031	0.404	0.429	0.040	0.031
Matrix type	Divisibility				Modularity			
	MEAN		STD		MEAN		STD	
	Solo	Joint	Solo	Joint	Solo	Joint	Solo	Joint
1. Cross-attention on Layer 0	0.691	0.683	0.018	0.014	0.076	0.066	0.068	0.065
2. Cross-attention on Layer 1	0.739	0.749	0.048	0.047	0.142	0.183	0.110	0.107
3. Average of cross-attentions	0.699	0.692	0.015	0.012	0.093	0.088	0.052	0.052
4. Element-wise product between cross-attentions	0.788	0.792	0.046	0.041	0.244	0.256	0.077	0.087
5. Self-attention on Layer 0	0.674	0.672	0.006	0.004	0.027	0.032	0.033	0.029
6. Self-attention on Layer 1	0.688	0.677	0.018	0.007	0.062	0.030	0.070	0.048
7. Average of self-attentions	0.683	0.676	0.012	0.005	0.050	0.034	0.052	0.033
8. Element-wise product between self-attentions	0.694	0.680	0.022	0.008	0.067	0.039	0.080	0.055
9. Average of all the attention on every layer	0.692	0.686	0.012	0.008	0.075	0.068	0.047	0.032
10. Element-wise product between all the attention on every layer	0.807	0.803	0.060	0.040	0.239	0.254	0.099	0.073
11. Element-wise product between the averages of cross and self-attention	0.713	0.700	0.020	0.015	0.108	0.106	0.059	0.057

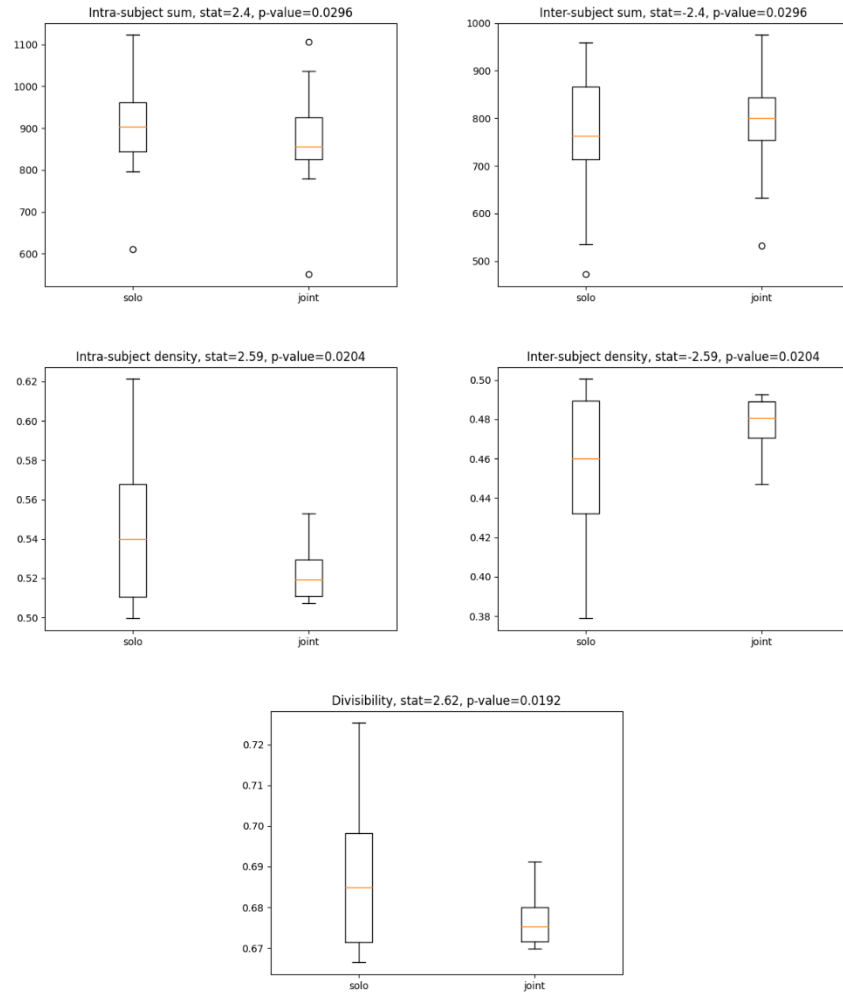


Figure 15: Box plots related to significances found for the related indices computed on “Self-attention on Layer 1” causality matrices.

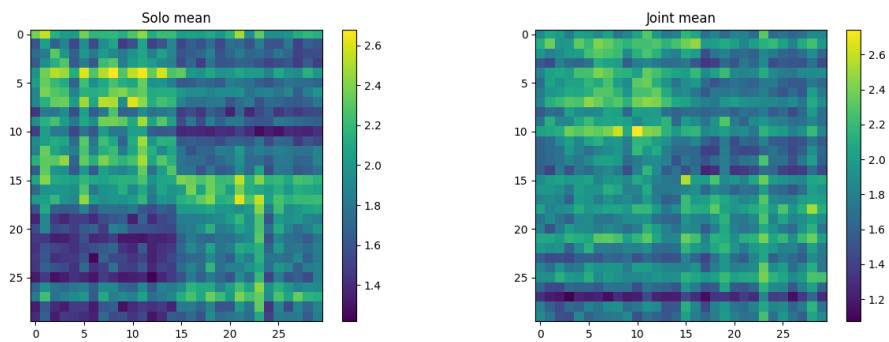


Figure 16: “Self-attention on Layer 1” causality matrices for the two conditions. “Solo condition” on the left, “Joint condition” on the right. The first half of the channels belong to subject A, the second half to subject B.

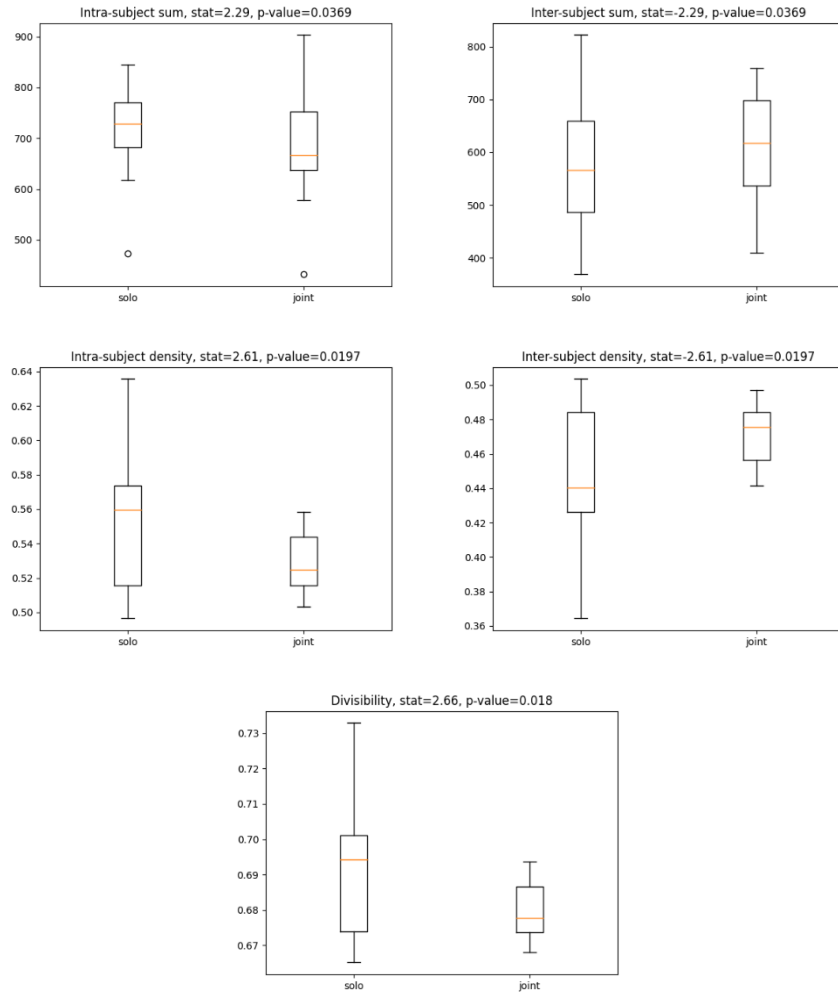


Figure 17: Box plots related to significances found for the related indices computed on “Element-wise product between self-attentions” causality matrices.

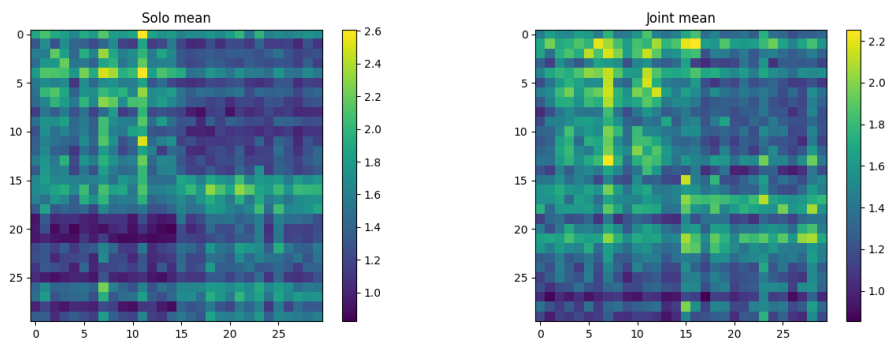


Figure 18: “Element-wise product between self-attentions” causality matrices for the two conditions. “Solo condition” on the left, “Joint condition” on the right. The first half of the channels belong to subject A, the second half to subject B.

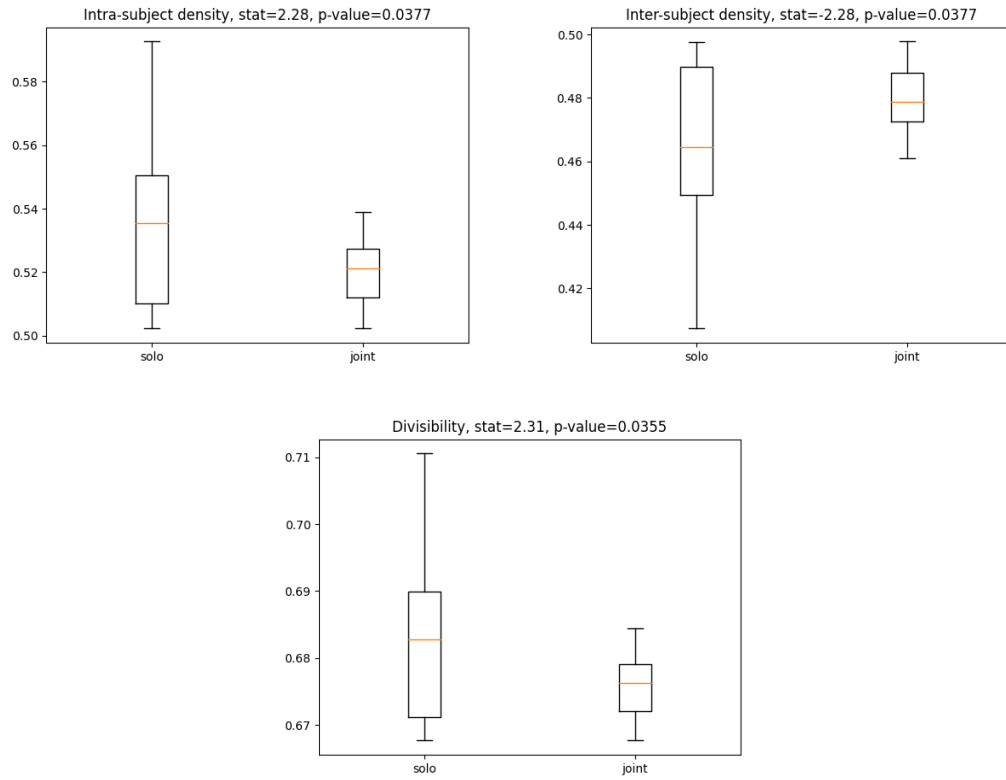


Figure 19: Box plots related to significances found for the related indices computed on “Average of self-attentions” causality matrices.

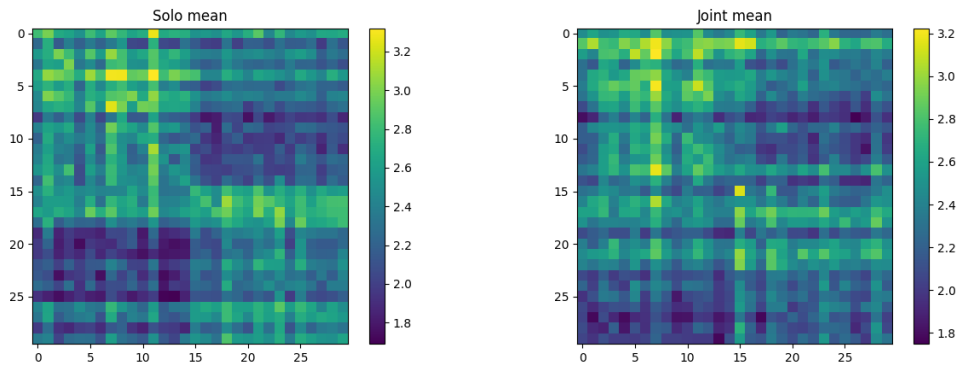


Figure 20: “Average of self-attentions” causality matrices for the two conditions. “Solo condition” on the left, “Joint condition” on the right. The first half of the channels belong to subject A, the second half to subject B.



# Conclusion and future works

My thesis work consisted of the design and experimental study of two novel non-linear methods as alternatives to traditional linear state-of-the-art approaches for the estimation of causality among brain signals. Specifically, I started with the idea to use the Spacetimeformer as an alternative model for the computation of Granger Causality, to replace the linear autoregressive models commonly used for the task. Later, I came to the second method to move the focus from the model residuals to the parameter that can be extracted from the attention matrices. This led to the application of an entirely new method in the field of neuroscience, and moreover to the recent field of hyperscanning EEG data, on which the application of non-linear methods is even more of scientific interest. To the best of my knowledge, this is the first time that Deep Learning methods are used for causality estimation in hyperscanning EEG data.

Among the two proposed methods, the second one seems to be the most promising, both because of its originality and because of the significantly reduced computational time compared to the use of the first method. In fact, it is based on a single general model, and it does not require to train as many restricted models as the number of channels in addition to the complete one. Specifically, in the experiments I conducted on hyperscanning data, the generation of a causality matrix using the first method takes about 30 times the time required by the second one (i.e., around 450-600 min vs. 15-20 min on a Tesla T4).

In addition, the second method provided physiologically meaningful and sound results, supported by statistical evidence.

For these reasons, I think that the attention-based method for the estimation of brain connectivity measures proposed in this thesis is worth to be explored in more depth.

Aspects that in my opinion would be worthwhile to be explored in the future are as follows:

- Experimenting with different aggregation functions. In fact, it has been proved by my experiments that the choice of the layer used to compute attention and the aggregation function used have a significant impact on the quality of the method, since statistical significances have been found only on a subset of matrix types.
- Another interesting experimentation to be carried out is about the number of decoder and encoder layers. The fact that only self-attention layers have shown to contain explainable information via attention matrices inspection is a hint that the number of layers could be a significant factor.
- In order to further evaluate the quality of the method, correlating computed indices with behavioural data would be a suitable test to do, to show if the properties of the connectivity matrices are related to the behavioural performances of the subjects (e.g., if dyads that perform better in the joint task also show more integrated multi-subject connectivity networks).

- Finally, it would be interesting to explore if the same attention method can be used for estimating temporal relationships between time samples in EEG data. In this case, while computing the attention matrices, instead of averaging over the temporal dimension, it would be appropriate to average over the spatial dimension.

# References

- Abhang, P. A., Gawali, B. W., & Mehrotra, S. C. (2016). Chapter 2 - Technological Basics of EEG Recording and Operation of Apparatus. In P. A. Abhang, B. W. Gawali, & S. C. Mehrotra (A cura di), *Introduction to EEG- and Speech-Based Emotion Recognition* (p. 19-50). Academic Press. doi:<https://doi.org/10.1016/B978-0-12-804490-2.00002-6>
- Analyzer, B. V. (2006). User manual. *Brain Products GmbH*.
- Anzolin, A., Toppi, J., Petti, M., Cincotti, F., & Astolfi, L. (2021). SEED-G: Simulated EEG Data Generator for Testing Connectivity Algorithms. *Sensors*, 21. doi:10.3390/s21113632
- Astolfi, L., Toppi, J., Ciaramidaro, A., Vogel, P., Freitag, C. M., & Siniatchkin, M. (2020). Raising the bar: Can dual scanning improve our understanding of joint action? *NeuroImage*, 216, 116813. doi:<https://doi.org/10.1016/j.neuroimage.2020.116813>
- Avanzolini, G., & Magosso, E. (2015). *Strumentazione biomedica: progetto e impiego dei sistemi di misura*. Pàtron.
- Babiloni, F., & Astolfi, L. (2014). Social neuroscience and hyperscanning techniques: past, present and future. *Neuroscience & Biobehavioral Reviews*, 44, 76–93.
- Baccala, L., & Sameshima, K. (2001, May). Partial directed coherence: A new concept in neural structure determination. *Biological Cybernetics*, 84, 463-474. doi:10.1007/PL00007990

- Bosga, J., & Meulenbroek, R. G. (2007). Joint-action coordination of redundant force contributions in a virtual lifting task. *Motor Control*, 11, 235–258.
- Cohen, M. X. (2014, January). *Analyzing Neural Time Series Data: Theory and Practice*. The MIT Press. doi:10.7551/mitpress/9609.001.0001
- Faes, A., Vantieghem, I., & Van Hulle, M. M. (2022). Neural Networks for Directed Connectivity Estimation in Source-Reconstructed EEG Data. *Applied Sciences*, 12. doi:10.3390/app12062889
- Grigsby, J., Wang, Z., & Qi, Y. (2021). Long-range transformers for dynamic spatiotemporal forecasting. *arXiv preprint arXiv:2109.12218*.
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., . . . others. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, 77, 354–377.
- He, B., Yuan, H., Meng, J., & Gao, S. (2020). Brain–computer interfaces. In *Neural engineering* (p. 131–183). Springer.
- Hsu, H., & Lachenbruch, P. A. (2014). Paired t test. *Wiley StatsRef: statistics reference online*.
- Ioffe, S. (2010). Improved consistent sampling, weighted minhash and l1 sketching. *2010 IEEE international conference on data mining*, (p. 246–255).

- Kaminski, M., & Blinowska, K. J. (2022). From Coherence to Multivariate Causal Estimators of EEG Connectivity. *Frontiers in Physiology*, 13. doi:10.3389/fphys.2022.868294
- Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29, 169–195.
- Marple Jr, S. L., & Carey, W. M. (1989). Digital spectral analysis with applications. *Digital spectral analysis with applications*. Acoustical Society of America.
- Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18, 544–551.
- Nakamura, K., Derbel, B., Won, K.-J., & Hong, B.-W. (2021). Learning-Rate Annealing Methods for Deep Neural Networks. *Electronics*, 10. doi:10.3390/electronics10162029
- Nicolas-Alonso, L. F., & Gomez-Gil, J. (2012). Brain computer interfaces, a review. *sensors*, 12, 1211–1279.
- Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452, 48–62.
- Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16, 051001.

- Schlögl, A. (2006). A comparison of multivariate autoregressive estimators. *Signal Processing*, 86, 2426-2429. doi:<https://doi.org/10.1016/j.sigpro.2005.11.007>
- Seth, A. (2007). Granger causality. *Scholarpedia*, 2, 1667. doi:10.4249/scholarpedia.1667
- Shojaie, A., & Fox, E. B. (2022). Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9, 289–319.
- Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE access*, 7, 53040–53065.
- Smith, S. (2022). An Easy Guide to Neuron Anatomy with Diagrams. *An Easy Guide to Neuron Anatomy with Diagrams*. Tratto da <https://www.healthline.com/health/neurons#anatomy>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wiener, N., Wiener, N., Mathematician, C., Wiener, N., Wiener, N., & Mathématicien, C. (1949). *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications* (Vol. 113). MIT press Cambridge, MA.
- Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, 31, 1235–1270.