From the Institute for Systems Neuroscience

at the Heinrich-Heine-University Düsseldorf

Interpretability and Reliability in Neuroimaging

Dissertation

to obtain the academic title of Doctor of Philosophy (Ph.D.) in Medical Sciences

from the Faculty of Medicine at Heinrich-Heine-University Düsseldorf

submitted by

Jean-Philippe Kröll

(2025)

As an inaugural dissertation printed with permission of the

Faculty of Medicine at Heinrich-Heine-University Düsseldorf

signed:

Dean: Prof. Dr. med. Nikolaj Klöcker

Examiners: PD Dr. Susanne Weis,

Prof. Dr. Julian Caspers

Parts of this work have been accepted and published:

Jean-Philippe Kröll, Simon B. Eickhoff, Felix Hoffstaedter, Kaustubh Patil, (2020). Evolving complex yet interpretable representations: application to Alzheimer's diagnosis and prognosis,*2020 IEEE Congress on Evolutionary Computation (CEC),* pp. 1-8. 10.1109/CEC48606.2020.9185843

Jean-Philippe Kröll, Patrick Friedrich, Xuan Li, Kaustubh R. Patil, Lisa Mochalski, Laura Waite, Xing Qian, Michael WL Chee, Juan Helen Zhou, Simon Eickhoff, Susanne Weis, (2023). Naturalistic viewing increases individual identifiability based on connectivity within functional brain networks. *NeuroImage*, *Vol*(273), https://doi.org/10.1016/j.neuroimage.2023.120083.

Mochalski, L. N., Friedrich, P., Li, X., Kröll, J.-P., Eickhoff, S. B., & Weis, S., I. (2024). Inter- and intra-subject similarity in network functional connectivity across a full narrative movie. *Human Brain Mapping, 45(11),* https://doi.org/10.1002/hbm.26802

Zusammenfassung

Die Entwicklung von Biomarkern auf der Grundlage der Magnetresonanztomographie (MRT) ist ein ständiges Bestreben auf dem Gebiet der klinischen Neurowissenschaften. Obwohl diese Biomarker ein großes Potenzial haben, wurden bisher nur wenige für den routinemäßigen klinischen Einsatz übernommen. Die größten Herausforderungen bei der Umsetzung in die klinische Anwendung sind die Genauigkeit, Zuverlässigkeit und Interpretierbarkeit eines Biomarkers. In dieser Dissertation wird daher ein neues maschinelles Lernverfahren (ML) vorgestellt, das die Genauigkeit der Diagnose und Prognose einer der häufigsten neurologischen Erkrankungen, der Alzheimer-Krankheit, durch die Konstruktion komplexer Darstellungen von Basis-Featuren verbessert. Durch die Verwendung einer kontextfreien Grammatik werden die konstruierten Repräsentationen gezwungen, menschlich interpretierbar zu bleiben, was die Validierung einer Beziehung zwischen dem Biomarker und dem vermuteten zugrunde liegenden pathologischen Korrelat ermöglicht. Darüber hinaus wird untersucht, ob *Naturalistic Viewing* (NV) Paradigmen geeignet sind, die für die Entwicklung von Biomarkern wichtigen Eigenschaften von MRT-Messungen zu verbessern, wie z. B. Reliabilität, geringere Variabilität innerhalb von Probanden und verbesserte Erkennung individueller Unterschiede im Vergleich zu Ruhemessungen (RS). Daher wird die Wirkung von NV-Stimuli mit unterschiedlichem sozialem Inhalt und unterschiedlicher Länge in 14 funktionellen Gehirnnetzwerken untersucht. Es wird gezeigt, dass NV-Stimuli, basierend auf der funktionellen Netzwerkkonnektivität (NFC), die Erkennung von individuellen Unterschieden in 10 von 14 Netzwerken verbessern, wobei die Stimuli mit dem höchsten Grad an sozialem Inhalt die größte Verbesserung erzielen. Eine anschließende Analyse bestätigt, dass Filmstimuli mit einem höheren Maß an sozialem Inhalt ähnliche NFC-Muster hervorrufen, die sich von RS und einem Stimulus ohne soziale Interaktionen unterscheiden. Darüber hinaus wird gezeigt, dass NV-Stimuli die Intra-Subjekt-Variabilität in meta-analytischen Netzwerken reduzieren können, die für die Wahrnehmung und Verarbeitung von Handlungen, Verhalten und Emotionen wichtig sind. Zusätzlich wird gezeigt, dass NV-Stimuli die Zuverlässigkeit von Graph-Metriken, die aus NFC extrahiert werden, gegenüber RS erhöhen können. Die Ergebnisse machen jedoch auch deutlich, dass NV-Stimuli die Metriken nicht uneingeschränkt über das gesamte Gehirn hinweg verbessern. Insbesondere für Netzwerke, die mit intrinsisch orientierten Funktionen verbunden sind, erweist sich RS als das zu bevorzugende Paradigma. Daher ist die Auswahl des geeigneten Stimulus und des funktionellen Netzwerks für die Beantwortung der jeweiligen Forschungsfrage von entscheidender Bedeutung. Schließlich stellt diese Dissertation einen neuen öffentlich zugänglichen NV-Datensatz zur Verfügung, um die Wirkung von NV-Stimuli weiter zu analysieren.

Summary

The development of magnetic resonance imaging (MRI) based biomarkers is a constant endeavor in the field of clinical neuroscience. Although these biomarkers hold great potential, only few have been adopted for routine clinical use. Primary challenges for the translation into clinical use are accuracy, reliability and interpretability of a given biomarker. Consequently, this dissertation presents a new machine learning (ML) framework that improves accuracy of diagnosis and prognosis of one of the most common neurological diseases, Alzheimer' Disease (AD), by constructing complex representations of base features. Further, by using a context-free grammar (CFG), the constructed representations are forced to remain humanly interpretable, thus enabling the validation of a relationship between the biomarker and the supposed underlying pathologic correlate. Additionally, it is investigated if naturalistic viewing (NV) paradigms are suited to improve characteristics of MRI measurements that are important for biomarker development, such as reliability, reduced intra-subject variability and enhanced detection of individual differences, in comparison with resting-state (RS). Therefore, the effect of NV stimuli with varying levels of social content and different lengths is investigated in 14 functional brain networks. It is shown that, based on network functional connectivity (NFC), NV stimuli improve the detection of individual differences in 10 out of 14 networks, with the stimuli with the highest level of social content achieving the most improvement. A subsequent analysis confirms that movie stimuli with higher levels of social content evoke similar NFC patterns that are distinct from RS and a stimulus lacking social interactions. Further, it is demonstrated that NV stimuli can reduce intra-subject variability in meta-analytic networks that are essential for perception and processing of action, behavior and emotions. In addition, it is shown that NV stimuli can increase the reliability of graph metrics extracted from NFC, over RS. However, the results also emphasize that NV stimuli do not unconditionally improve metrics of interest across the whole brain. In particular for networks that are related to intrinsically oriented functions, RS proves to be the more favorable paradigm. Therefore, selecting the appropriate stimulus and functional network is essential for addressing the specific research question at hand. Finally, this dissertation provides a new publicly available NV dataset to further analyze the effect of NV stimuli.

List of abbreviations

Abbreviation	Definition
MRI	magnetic resonance imaging
AD	Alzheimer's Disease
ML	machine learning
fMRI	functional magnetic resonance imaging
sMRI	structural magnetic resonance imaging
rs-fMRI	resting-state functional magnetic resonance imaging
NV	naturalistic viewing
AI	artificial intelligence
SVM	support vector machine
RBF	radial basis function
GE	grammatical evolution
FC	functional connectivity

RS	resting state
NFC	network functional connectivity
AM	autobiographical memory
ER	emotion regulation
SM	semantic memory
ToM	theory of mind
eSAD	extended socio-affective default
FNM	full narrative movie
ADHD	attention-deficit/hyperactivity disorder

Table of contents

1 Introduction

Since initial discoveries that MRI-measured structural brain differences between healthy individuals and patients can be used to monitor or even predict disease progression, many researchers have tried to develop biomarkers for use in clinical settings. Apart from conventional statistical approaches, the use of ML based methods has gained a lot of popularity. Especially with more availability of larger datasets, researchers have turned to ML as such methods better handle complex, high-dimensional data than conventional approaches. However, one of the primary challenges in the translational use of ML methods is the lack of explainability, particularly with non-linear techniques. Explainable (i.e., human-interpretable) methods on the other hand not only offer valuable insights into disease mechanisms but also foster clinician-patient trust, which is crucial for the broader social acceptance of ML approaches. Besides the need for interpretable ML models, the reliability of MRI measurements itself are an important factor for ensuring accurate and reproducible results. In functional MRI (fMRI) research, resting-state fMRI (rs-fMRI) has been the gold standard for the study of brain connectivity because it measures intrinsic functional organization independent of task constraints. However, rs-fMRI is not without limitations, including intra-subject variability, susceptibility to motion artifacts, and the influence of unconstrained mental processes. All these compromise the reliability of rs-fMRI. NV paradigms, where participants engage with dynamic, real-world stimuli, offer a promising alternative to traditional rs-fMRI for studying brain connectivity. However, the reliability of NV paradigms and their ability to capture individual differences not only across the brain, but also in functional networks, are yet to be assessed.

1.1 Alzheimer's Disease and Interpretability

One of the diseases which has extensively been studied with MRI measures is AD. AD is the most common form of dementia, affecting about 50 million people worldwide. It significantly impairs memory, language, and intellectual capabilities, making day-to-day tasks increasingly difficult for patients. With increasing life expectancy, AD has become an emerging public health problem (Nandi et al., 2022). It has therefore become a main research objective to develop accurate methods for early diagnosis of AD. One promising method of distinguishing AD from healthy controls is the application of machine learning to structural MRI (sMRI) data (Lahmiri and Shmuel, 2019; Zhu et al., 2021). Since brain atrophy is a feature of AD and can

be observed in sMRI scans, most machine learning approaches have been successful in leveraging this information for classification. Since sMRI is already a staple in clinical practice, developing an accurate diagnostic tool based on such scans is of great clinical value. However, aside from accuracy, another critical factor is interpretability because it enables researchers to understand which areas of the brain are most affected by AD and how these regions interact. That kind of insight can open up knowledge on mechanisms of the disease and ultimately help improve treatment outcomes. Furthermore, building trust between clinicians using ML/artificial intelligence (AI) methods and patients is essential for the acceptance of these technologies. This can be achieved by clearly explaining the reasoning behind decisions and the uncertainties associated with different options. Therefore, the latest EU guidelines for trustworthy AI, make transparency one of the main requirements for the application of machine learning algorithms ("EU guidelines on ethics in artificial intelligence: Context and implementation," 2019). Accuracy and interpretability, however, usually come with a trade-off. While interpretable models like linear support vector machine (SVM) and logistic regression can not capture very complex feature interactions, more complex models like SVM with radial basis function (RBF) kernel or neural networks would be capable of recognizing such patterns, but they are less interpretable in terms of how they arrived at those decisions.

One approach that can be utilized to enhance both accuracy and interpretability is employing grammatical evolution (GE) for feature construction and selection. GE is an evolutionary algorithm which combines the concepts of genetic programming and formal grammar systems with the aim of evolving solutions to complex problems. Like all evolutionary algorithms, GE works by maintaining a population of solutions (i.e. newly constructed features), selecting the fittest individuals for reproduction, and applying genetic operators like crossover and mutation to create new candidates. Each individual is evaluated based on a fitness function, and over multiple generations, the population evolves toward better solutions. However, unlike other evolutionary algorithms where program structures are evolved explicitly, GE uses a user-defined formal grammar. By restricting this grammar to perform only basic arithmetic operations during feature generation, interpretability of the resulting features can be enforced.

1.2 Resting-State vs Naturalistic Viewing

Most research on functional connectivity (FC) has focused on connectivity patterns observed during task-free resting state (RS), where participants lie in a scanner without engaging in any specific task (Amft et al., 2015; Damoiseaux et al., 2006; Langner and

Eickhoff, 2013). RS is believed to reflect the brain's intrinsic organization and has also been shown to align well with findings derived from task-based studies (Smith et al., 2009). Further, the ease of implementing RS data allows for the rapid acquisition of large healthy and clinical samples due to minimal participant demands. While the RS paradigm has provided valuable insights into brain organization, it also has limitations: RS data can be heavily influenced by head movement and drowsiness due to its unconstrained nature (Tagliazucchi and Laufs, 2014; Van Dijk et al., 2012), as participants struggle to stay awake and still without a task or stimulus. Moreover, RS is susceptible to the influence of spontaneous thoughts (Christoff et al., 2004; Gonzalez-Castillo et al., 2021).

NV paradigms, where participants watch a story or film, have recently gained popularity as they offer a more ecologically valid approach to studying brain function. Compared to RS, NV offers several advantages. By providing a stimulus, NV reduces the variability caused by spontaneous thoughts. Furthermore, NV has been shown to reduce fatigue and head movement by increasing participant engagement, as compared to RS (Finn and Bandettini, 2020; Vanderwal et al., 2019). Finally, watching movies can make scanning more tolerable for participants who find it challenging to stay still (e.g., individuals with ADHD) or complete demanding tasks (e.g., those with cognitive impairments) (Eickhoff et al., 2020).

1.3 The two present Samples

Since the rise of NV, a plethora of samples have been made available. Researchers have implemented different stimuli, varying from short clips that last less than two minutes to full length movies (DuPre et al., 2020). More so, movie clips differ in their content. On the one hand, studies have used rather neutral clips e.g. depicting landscapes or even more abstract clips like the movie Inscapes as a baseline comparison to RS (Van Essen et al., 2012; Vanderwal et al., 2015). On the other hand, many authors have suggested that movie clips with social content are more likely to engage participants (Finn et al., 2018; Finn and Bandettini, 2020; Nguyen et al., 2019; Rikandi et al., 2017). Related, several studies have suggested that the cultural background of a person can influence the effect of NV. Cultural norms, values, and experiences shape one's interpretation of social interactions and narrative elements depicted in movie stimuli. Consequently, elucidating the interplay between cultural background and naturalistic viewing is crucial for understanding the variability in neural responses across populations (Eickhoff et al., 2020).

To address these points, two identical samples were acquired for this dissertation. Both samples employed the same three different movie stimuli with different levels of social content. The first movie was the movie Inscapes which depicts only abstract animations and lacks any form of social interaction. The second movie, The Circus (United Artists Digital Studios, 1928, directed by Charlie Chaplin) is a silent black-and-white film that shows the protagonist being chased through a circus by the police and unintentionally causing comic situations during his escape. Due to the lack of spoken words and the chaplin-typical pantomime-esque depiction, this movie is employed as a stimulus with moderate level of social interactions. The third movie, Indiana Jones and the Temple of Doom (Paramount Pictures, 1984, directed by Steven Spielberg) the protagonist is shown during an intense negotiation and afterwards has to fend off several hitmen who try to kill him. Due to the complex interactions between the characters during this scene, the movie is seen as the stimulus with the highest level of social content. To enable the comparison of cultural effects, one sample was acquired in Singapore and one sample in Jülich.

1.4 Meta-analytic Networks

The human brain is commonly seen as being organized into modules of spatially distinct areas that form functional networks (Sporns and Betzel, 2016). These networks correspond to particular cognitive domains, such as memory (Spreng et al., 2009), social cognition (Bzdok et al., 2012) and executive function (Rottschy et al., 2012). Since NV paradigms use complex, multimodal stimuli that elicit activation patterns across the whole brain, adopting a network perspective can explain the effect of movie stimuli on particular cognitive processes.

In the context of NV, one would expect that networks that relate to different functions, should also differ in their response to the same stimulus. For example, a functional network that processes emotions should be differently affected by a movie scene with strong emotional content, in comparison to the motor network. Therefore, investigating FC in networks that cover different cognitive domains under NV, extends the knowledge over traditional whole-brain studies. There are various methods to define functional networks (Power et al., 2011; Schaefer et al., 2018; Smith et al., 2009) one of which is the use of meta-analysis (Eickhoff et al., 2012). Meta-analytically defined networks integrate converging data from a multitude of studies and thus represent the most likely core nodes that are involved in a given function. Therefore, studying FC in meta-analytical networks could offer new and more detailed insights into the effects of naturalistic viewing, as compared to conventional whole-brain studies.

1.5 Individual Differences

While traditional neuroscience has mostly focused on group-level analysis, exploring variability between subjects is essential for a comprehensive understanding of individual brain function. Characterizing individual variations in FC offers additional insights into the relation of brain function, behavior and cognition. Furthermore, individual differences hold significant implications for personalized medicine. Understanding variations between individual brains will help to assess personal susceptibility to neurological disorders and response to interventions. However, the detection of individual differences in FC has been a challenging task. Due to motion artifacts and physiological fluctuations inherent to fMRI data, true individual differences are partly obscured and difficult to disentangle from noise (Dubois and Adolphs, 2016). In addition, the typically used RS paradigm is influenced by attention fluctuations and spontaneous thoughts of the participant (Christoff et al., 2004). Moreover, the passive nature of the RS paradigm might not fully capture the individual's cognitive abilities, thus limiting the sensitivity with which subtle variations across participants can be detected. Previous research has shown that certain tasks improve the sensitivity to individual differences in FC in comparison to RS (Finn et al., 2017). However, the authors themselves point out that alternative paradigms are worth exploring for the analysis of individual differences, as they might enhance the detection of individual differences over RS and task approaches. One of these paradigms is NV. In contrast to RS and task, NV employs rich stimuli that better reflect the complexity of real-life experiences. By exposing participants to a wide array of sensory, emotional and contextual input, NV stimuli probe the human brain under a condition that allows past experiences, cultural beliefs and cognitive strategies to shape the neuronal response. Thereby, NV imposes richer brain state dynamics and therefore more individual connectivity profiles, which might better reflect individual characteristics than RS (Vanderwal et al., 2017).

1.6 Reliability

fMRI has become an indispensable tool in neuroscience research and has granted substantial insight into the function of the human brain. As applications of fMRI expand to the prediction of clinical outcomes, the reliability of the measurement has become a major concern. In order to guide clinicians in the diagnosis and prognosis of brain disorders, a measure has to consistently give accurate results. However, the reported reliability of fMRI measures varies vastly across studies (Bennett and Miller, 2010), partly due to small test-retest samples, but also due to different analysis choices. Therefore, finding methods that increase reliability has

become a priority. Traditionally, the field has relied either on task-free RS or on highly controlled task designs. Although both paradigms have their benefits, the former lacks specificity which makes it challenging to relate the observed neural activity to function, while the latter has limited generalizability because it uses highly artificial tasks to focus on one specific cognitive process. One of the methods with potential to increase reliability of fMRI is NV, because it engages the brain in a more structured, yet ecologically valid context. NV paradigms present participants with stimuli that mimic conditions under which the brain naturally operates, such as movies depicting dynamic social interactions. Thereby, participants might react in a manner that is more reflective of their typical cognitive processes, possibly leading to more consistent and reliable results.

1.7 Ethics protocols

The acquisition and use of the JUMAX dataset has been approved by the Heinrich-Heine-University Düsseldorf (Study-Nr. 2019-791). The IMAX dataset was acquired under protocols approved by the National University of Singapore (NUS-IRB REFERENCE CODE: B-14-045). Data collection and sharing for the ADNI project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904).

1.8 Aims of the thesis

This thesis aims to advance the development of biomarkers. Therefore, a new ML framework is provided that can accurately diagnose and prognose AD while retaining interpretability of the model. Interpretability is one of the major hurdles for the translation of ML based MRI research to real-world applications. Subsequently, the thesis will focus on another key challenge of biomarker development which is improving the reliability and the ability to detect individual differences. The thesis will explore the use of NV paradigms as an alternative to resting-state fMRI (rs-fMRI) for studying brain connectivity. It will assess how NV paradigms capture individual differences in functional networks and how different stimuli influence within- and between-subject similarity. In addition, the reliability of NV will be analyzed and compared to that of RS.

This dissertation pertains to four studies. Study 1 establishes a ML framework that maximizes predictive accuracy while retaining feature interpretability. The framework is applied to the diagnosis and prognosis of AD. The features of the final ML model are then examined in terms of their interpretability. Study 2 investigates the effect of NV on FC in fourteen meta-analytic

networks. Particularly, the study focuses on the identifiability of individuals based on FC during NV and RS. In addition, individual variability in network FC (NFC) is assessed by comparing within- and between-subject similarity during NV and RS. These results are then compared between different NV stimuli and functional networks. Study 3 investigates the within- and between-subject similarity in NFC of the same 14 networks using a full narrative movie (FNM) and employing a linear mixed model to assess which factors explain inter- and intra-subject similarity. Study 4 focuses on the reliability of NV and compares it to that of RS. The influence of NV on reliability is characterized on the basis of graph metrics extracted from the same 14 functional networks.

2 Evolving complex yet interpretable representations: application to Alzheimer's diagnosis and prognosis, Jean-Philippe Kröll, Simon B. Eickhoff, Felix Hoffstaedter, Kaustubh Patil, *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-8, (2020)

Evolving complex yet interpretable representations: application to Alzheimer's diagnosis and prognosis

Jean-Philippe Kröll
Inst. of Neurosci. and Medicine, INM-7, Forschungszentrum Jülich, and Inst. of Systems Neurosci., HHU Düsseldorf
Germany
j.kroell@fz-juelich.de

Simon B. Eickhoff
Inst. of Neurosci. and Medicine, INM-7, Forschungszentrum Jülich, and Inst. of Systems Neurosci., HHU Düsseldorf
Germany
s.eickhoff@fz-juelich.de

Felix Hoffstaedter
Inst. of Neurosci. and Medicine, INM-7, Forschungszentrum Jülich, and Inst. of Systems Neurosci., HHU Düsseldorf
Germany
f.hoffstaedter@fz-juelich.de

Kaustubh R. Patil
Inst. of Neurosci. and Medicine, INM-7, Forschungszentrum Jülich, and Inst. of Systems Neurosci., HHU Düsseldorf
Germany
k.patil@fz-juelich.de

***Abstract*—With increasing accuracy and availability of more data, the potential of using machine learning (ML) methods in medical and clinical applications has gained considerable interest. However, the main hurdle in translational use of ML methods is the lack of explainability, especially when non-linear methods are used. Explainable (i.e. human-interpretable) methods can provide insights into disease mechanisms but can equally importantly promote clinician-patient trust, in turn helping wider social acceptance of ML methods. Here, we empirically test a method to engineer complex, yet interpretable, representations of base features via evolution of context-free grammar (CFG). We show that together with a simple ML algorithm evolved features provide higher accuracy on several benchmark datasets and then apply it to a real word problem of diagnosing Alzheimer's disease (AD) based on magnetic resonance imaging (MRI) data. We further demonstrate high performance on a hold-out dataset for the prognosis of AD.**

Keywords — grammar evolution, feature representation, interpretability, Alzheimer's disease, machine learning

I. Introduction

Application of machine learning and artificial intelligence (AI) methods in medical and clinical problems has gained increasing attention in recent years [1][2]. These methods can find patterns in high-dimensional data and thus have the potential to provide gains in diagnostic and prognostic accuracy. However, there are also skepticisms and societal concerns, especially regarding the explainability of the models and their predictions [1][3][4]. According to the latest EU guidelines for trustworthy AI, transparency is one of the main requirements for the application of machine learning algorithms [5]. Importantly, fostering trust between clinicians assisted by ML/AI methods and patients by communicating reasons behind decisions and uncertainties associated with options is crucial for the acceptance of ML methods [6], [7]. In addition, explainable/human-interpretable models are inherently beneficial in a clinical setting as they can help understand the biology underlying disease mechanisms and disease progression. It is, therefore, important to develop methods and frameworks that can simultaneously provide high accuracy and interpretability.

Feature engineering is one of the key concepts to improve model performance: Processing the available features in such a way that they are easily learnable by a classifier is arguably one of the most important parts of machine learning [8]. Single features may seem irrelevant until considered in combination with others. Often, exhaustively exploring the complete range of possible feature combinations is computationally too expensive, due to the high dimensionality of the data. Evolutionary algorithms can improve the search in such combinatorial problems by systematically searching the space guided by the usefulness of the candidate solutions. Previous work utilizing evolutionary algorithms have shown promise in various research areas. Some of these approaches have relied on grammatical evolution (GE) [9] for feature selection and generation. For example, Silva et al. employed GE to select and generate features for the prediction of the daily peak electricity load in planning of power systems [10]. Implementing a combination of GE and neural networks, Gavrilis et al. generated new features and could thereby improve performance on nine out of ten classification datasets [11]. Demonstrating its suitability for medical purposes, Smart et al. similarly utilized GE to select the best subset of features as well as to generate new features for detecting epileptic oscillations in patients with epileptic seizures [12]. Motsinger et al. proposed a combination of GE and neural networks to perform automatic feature selection in genetic epidemiology [13]. In a study by Georgulas et al., GE was utilized to improve the classification of pathological fetal heart rate where artificial features were derived from the 19 original features and used to train a neural network [14]. These studies show that the models generally benefited from the constructed features (CF). If the generated features and the model are restricted to retain a human-interpretable form, such a feature generation framework can be leveraged to promote both accuracy as well as interpretability. Towards this goal, we propose a framework based on GE, which achieves a good trade-off between these two goals.

Building upon its promise in engineering new and useful features, here we use GE to evolve new feature representations as combinations of the original or base features which are then used as a basis for classification. Our motivation for using GE was to test a feature construction method that can produce human-interpretable features that meet the requirements for trustworthy AI. Although there exist other feature extraction/construction methods (e.g. PCA) the resulting features are often not interpretable.GE can limit the search space and efficiently construct new features by incorporating domain-specific knowledge and user expectations through a pre-defined

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

set of rules, the so-called 'grammar'. By restricting the grammar to basic arithmetic operations, we enforce the expectation on the engineered features to be human-interpretable. We then use the naïve Bayes (NB) classifier as a model. We first demonstrate utility of evolved feature representation on eight benchmark datasets. We then apply our framework to the clinical problem of diagnosis of the Alzheimer's disease —i.e. AD versus healthy control (HC) classification—using base features derived from structural MRI (sMRI) data. We expected our approach to generate human-interpretable features which include information about the interactions between brain regions. Additionally, we apply the AD vs. HC model to a hold-out set to probe its prognostic capacity—i.e. to predict whether a person with mild cognitive impairment (MCI) will develop AD or not.

Taken together, the main contributions of our work are: (1) we propose a GE framework to construct arithmetic combinations of base features which improves accuracy; and (2) by applying it to the real-world clinical problems of diagnosis and prognosis of AD, we demonstrate that the proposed framework can uncover complex yet interpretable interactions between brain regions.

This paper is structured as follows: Section II lays out the background of AD and briefly showcases current ML-based diagnostic approaches. Section III gives a brief introduction to GE and the general workflow. Section IV gives a detailed description of the feature construction method. In section V, the results are presented and discussed. Section VI presents the conclusions of our work.

II. Alzheimer's Disease and its Diagnosis and Prognosis

Among the estimated 50 million people suffering from dementia worldwide, AD is the most common form [15]. Disturbances in memory, language and higher executive functions lead to severe obstruction of a patient's life. With high prevalence in the elderly, AD has become a major public health problem, due to the increasing life expectancy of the population. It is, therefore, important to develop accurate and interpretable methods for early diagnosis of AD. One approach which has shown a good diagnostic promise—i.e. AD versus HC classification—is using sMRI derived features in combination with machine learning algorithms. Since the progression of AD is highly associated with loss of brain volume detectable in sMRI images, various algorithms capitalizing on atrophy in AD patients have shown good classification accuracy. Furthermore, as sMRI is routinely acquired in many clinics, a highly accurate and interpretable method using sMRI data has a high translational potential.

Utilizing support vector machine (SVM), Klöppel et al. classified grey matter segments of 20 pathologically proven AD patients and matched healthy controls with 96% accuracy [16]. On a larger dataset of 652 subjects, Liu et al. employed an ensemble method, based on sparse representation-based classifiers with an accuracy of 91% [17]. Lebedev et al. proposed random forest based ensembles and were able to differentiate AD from HC with an accuracy of 90% [18]. All of these approaches rely either on whole-brain analysis or atlas derived features. In most cases, classification is based on grey matter volumes of individual brain regions and benefits from the fact that areas highly affected in AD, like the hippocampus, provide good discrimination. In addition to high accuracy, it is desirable to have interpretable models that can help uncover the brain regions involved in AD and interactions between them. This, in turn, can help understand the disease mechanisms and progression leading to better treatment and care. However, a trade-off exists between these two goals such that the implicitly interpretable methods (e.g. linear SVM or logistic regression) do not implicitly take complex interactions between features into account, while other models do so with reduced implicit interpretability (e.g. RBF kernel SVM and neural networks).

III. Grammatical Evolution Framework

We consider the binary supervised learning problem where given a labeled dataset $D = \{(x_i, y_i)\}_{i=1}^{n}, x \in \mathbb{R}^d, y \in \{0,1\}$, we want to learn a mapping function $f: x \to y$ such that f generalizes on unseen data. Here, we use GE to evolve feature combinations using CFG to learn $f': x' \to y$, where $x' = \mathrm{CFG}(x)$, $x' \in \mathbb{R}^p, p \leq d$. Our aim is to identify f' such that it performs better than f and is still interpretable. We propose to use grammatical evolution for this.

Fig. 1 shows the general workflow of the proposed method. The initial population of chromosomes is translated into expressions using the production rules of the CFG. Note that each chromosome can result in a different number of expressions. Subsequently, new features are constructed by combining the base features according to the defined expressions (see section IV or details). The constructed features are used to train and evaluate a classifier in a cross-validated (CV) fashion to estimate generalization performance [19].

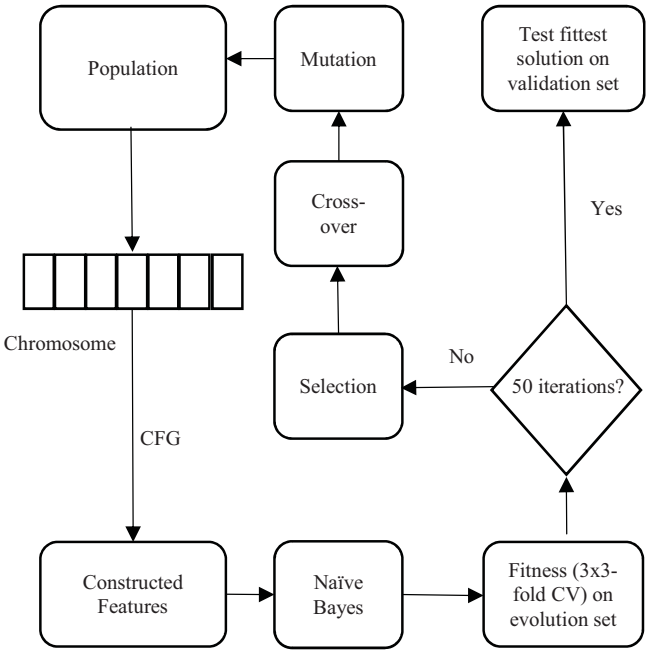


Fig.1 Workflow of the proposed framework

In CV, the data is randomly split into k equally sized subsets. One subset is retained as validation data, while the other k-1

folds are used to train the model. Consequently, each of the subsets becomes the validation data once, resulting in k different estimations. The mean of those results gives an estimate of how well the model will generalize on new data. Chromosomes then undergo selection, cross-over and mutation and are evolved to maximize their fitness. Even with the use of CV, studies have shown that optimization based solutions are prone to over-fit [20], [21]. To assess the true generalization performance of the constructed features, using CV is not enough as it is a part of the optimization process and can thus lead to overly optimistic estimates [22]. It is, therefore, necessary to assess the generalization performance outside the optimization procedure. To achieve this, the data is randomly split into two sets, 80% evolution set and 20% validation set. The evolution set is used to evaluate the features constructed by GE expressions during evolution. This is done using 3-times 3-fold CV. The fittest solution is then evaluated on the hold-out 20% validation set.

We selected NB as it provides two desirable properties, 1) its low variance and high bias which makes it less prone to over-fitting and therefore counteracts the susceptibility to overfitting within the GE optimization iterations, and 2) its probabilistic output which makes the predictions easier to communicate. Furthermore, NB is negatively affected by redundant and irrelevant features [23], so we expect the evolved feature representations to mostly contain relevant and non-redundant features.

We use the Brier score as cost (negative fitness) measure. Brier score was chosen as it is a proper scoring rule and hence can be used to rank solutions.

A. Grammatical Evolution

In this study, the gramEvol R-package (https://github.com/fnoorian/gramEvol) was used [24]. GE combines CFG and genetic algorithms to optimize programs towards a specific task. CFG is used to generate patterns of strings according to a set of recursive rules. The notation technique used here is the Backus-Naur form (BNF). The CFG is described by the tuple (T,N,R,S), where T is the set of terminal symbols, N is the set of non-terminals with $N \cap T = \emptyset$, R the set of production rules and S the start symbol, $S \in N$. Non-terminal symbols can be replaced by other non-terminal or terminal symbols whereas terminal symbols are literals. N and T together build the lexical elements which are used in the production rules R. R is defined as relations in the form of $x \rightarrow \alpha$ with $x \in N$, $\alpha \in (N \cup T)$. The user-defined grammar is utilized to impose a set of grammatical production rules which determine the chromosomes. Each gene denotes a production rule of the CFG. Following the predefined set of rules, genotypic integer strings are translated into functional phenotypic programs, a process which is called genotype-to-phenotype mapping. The mapping function is the mod rule, defined as:

$$R = B \bmod RN$$

Where B is the codon integer value, mod is the modulus operator and RN is the number of rules for the current non-terminal. Mapping begins at S and subsequently replaces each non-terminal element N, according to the production rule determined by the mapping function. Mapping continues until every non-terminal element is replaced by a terminal. If the chromosome runs out of codons before a valid expression could be produced, wrapping is applied. By reusing the codons, the mapping process continues. To prevent infinite recursions, wrapping is limited to a certain number and will result in a poor fitness score if the limit is reached. Details of the settings that we used and feature construction are provided in the next section.

The evolution was performed with a genetic algorithm (GA) [25]. GA is an optimization algorithm inspired by evolution in which generations of chromosomes, representing the genotype—i.e., candidate solutions, are successively optimized and evaluated based on a fitness measure. The chromosomes are then subject to selection, cross-over and mutation, producing the next generation. This process is repeated until terminal criteria like a certain threshold of fitness or the predefined number of generations are reached.

IV. Experimental Setup

The production rules of the CFG were defined as the grammar shown in Table I. Non-terminal symbols are expression, operator and variable and are enclosed by angle brackets. On the other side, terminals are the actual mathematical operators and original features. Thereby, the resulting expressions are arithmetic combinations of the original features. The feature construction process takes the values of each chromosome and applies the CFG rules from Table I to the base features (Tab. IIB).

To reduce computational cost whilst preserving the diversity of the solutions, population size was set to 20 chromosomes. Additionally, the number of generated features was fixed to be equal to or less than the number of original features of the given dataset, with 14 codons per expression. The mutation chance for each codon was set to $1/(\text{genomeLength}+1)$ and single-point cross-over was used. The initial population included the base model (all original features by themselves). Other chromosomes in the initial population were randomly created in the range of [0, d-1]. Evolution was terminated after 50 generations.

The cost (negative fitness) of each chromosome was calculated as stratified Brier score [26] to take the imbalanced nature of some datasets into account. Using the constructed features, an NB model was fit to the two training folds within the evolution set and used to predict the held-out fold. The predicted assignment probabilities were used to calculate the Brier score for each class separately. The two Brier scores were then averaged to get the cost value, with lower values indicating better performance. The settings of the GE are shown in Tab. IIA.

The optimized *GE model* using the 80% evolution set —i.e. the NB model on the constructed features—was evaluated on the 20% validation set. The same evolution set was used to build a *base model* using the original features and evaluated on the validation set. To consider the randomness in the evolution set-validation set split and the GE initialization, we ran the GE framework five times for each dataset. Four evaluation metrics are reported: area under the ROC curve (ROC), balanced accuracy (Acc), F1-score (F1) and stratified Brier score (Brier).

TABLE I. *GRAMMAR USED*

Rule			Rule number
S	::=	<expr>	0
<expr>	::=	<expr> <op> <expr>	0
		<var>	1
<op>	::=	+ | - | * | /	0|1|2|3
<var>	::=	X_1 | X_2 | ... | X_n	0|1|...|n-1

TABLE II. FEATURE CONSTRUCTION

A) SETTINGS OF THE GE

Parameters	Value
Number of individuals	20
Number of generations	50
Chromosome length	[0, d-1]
Mutation rate	1/(d+1)

B) EXAMPLE FEATURE CONSTRUCTION

String	Chromosome	Operation
<expr>	8,9,14,3,6,11,7,6,13,4	8 mod 2 = 0
<expr> <op> <expr>	9,14,3,6,11,7,6,13,4	9 mod 2 = 1
<var> <op> <expr>	14,3,6,11,7,6,13,4	14 mod 14 = 0
X_1 <op> <expr>	3,6,11,7,6,13,4	3 mod 4 = 3
X_1 * <expr>	6,11,7,5,13,4	6 mod 2 = 0
X_1 * <expr> <op><expr>	11,7,5,13,4	11 mod 2 = 1
X_1 * <var> <op> <expr>	7,5,13,4	7 mod 14 = 7
X_1 * X_8 <op> <expr>	5,13,4	5 mod 4 = 1
X_1 * X_8 + <expr>	13,4	13 mod 2 = 1
X_1 * X_8 + <var>	4	4 mod 14 = 4
X_1 * X_8 + X_5	**constructed feature**	

The original chromosome is [8,9,14,3,6,11,7,6,13,4]. The process starts with the first integer of the chromosome, in this case, eight. Since the start symbol is <expr>, which has two different rules, the first operation is 8 mod 2 = 0. Consequently, rule number 0 is selected and <expr> is translated into <expr> <op> <expr>. After that, the leftmost non-terminal is selected and the next integer is used to determine the following rule. The process is repeated until every non-terminal element is substituted by a terminal. The final expression is 'X_1 * X_8+ X_5'.

A. Datasets

We used eight real-world benchmark datasets from UCI (http://www.wisostat.uni-koeln.de/de/forschung/software-und-daten/data-for-classification/) and two real-world clinical datasets.

1) Breast Cancer Wisconsin: The sample contains 569 patients with breast cancer. The objective is to differentiate malignant and benign cases using 30 features computed from a fine needle aspirate of a breast mass, describing characteristics of the cell nuclei of the image. The database contains 257 benign and 212 malignant cases.

2) Pima Indians diabetes: This dataset contains 768 females of Pima Indian heritage. The objective is to predict diabetic status using eight diagnostic measurements. Variables include the number of pregnancies, glucose concentration in plasma, blood pressure, skin thickness, insulin concentration, BMI, age and Diabetes Pedigree Function. 268 of the subjects are diagnosed as diabetics.

3) Heart Disease: The sample contains 270 participants with 120 patients with diagnosed heart disease. The objective is to classify the absence or presence of heart disease using 13 features with various diagnostic measurements.

4) Irish: The dataset contains 500 instances of Irish school children. The objective here is to classify into male and female, based on five features dealing with the educational status of the children.

5) Image Segmentation: The dataset contains 660 outdoor images. The images were hand segmented to create a classification for every pixel. In this case, images are classified into "containing window" and "containing cement". 330 examples are available for each class.

6) Tennis: The dataset contains 87 instances of subjects under pain medication. Based on 15 features dealing with experienced drug efficacy, the objective is to classify into male and female.

7) Diabetes: The dataset contains 112 instances of diabetics. The objective is to differentiate the diabetic type based on five metabolic variables.

8) Crabs: The dataset contains 200 instances of Leptograpsus crabs. Based on 5 features describing physical attributes, the objective is to classify into male and female.

9 and 10) Alzheimer's Disease Neuroimaging Initiative (ADNI). We derived two datasets from the ADNI database [27]. (A) The AD diagnosis dataset contains 459 subjects with 3T scans with the objective to classify them as AD or HC. Structural (T1-weighted) MRI images of 153 AD patients and 306 HC are extracted. Utilizing the CAT toolbox (http://dbm.neuro.uni-jena.de/cat), voxel-based morphometry (VBM) is performed to estimate local grey matter volume. Subsequently, a brain atlas is applied which partitions the brain into 173 parcels. The brain atlas contains 100 Schaefer atlas parcels covering the cortex [28], complemented by 36 subcortical regions from Brainnetome [29] and 36 cerebellum parcels from Buckner et al [30]. The average grey matter volume within each of the 173 parcels is calculated as base features for each subject. (B) The MCI to AD prognosis dataset contains similarly derived 173 features for 267 subjects of which 138 later converted to AD. The objective here is to classify converters and non-converters.

V. EXPERIMENTAL RESULTS

In this section, we discuss the results obtained on the benchmark datasets as well as the two clinical datasets. Overall, we observed that both the GE constructed features as well as the base features combined with the naïve Bayes classifier were able

to classify most datasets with high accuracy. Notably, feature construction via GE resulted in superior performance in most cases.

A. Benchmark UCI Datasets

In Table III, the results from the application of both base and GE models on several test datasets are listed. We observed that for six out of eight datasets the NB model using GE constructed features outperformed the NB model using base features in at least four out of five runs on the stratified Brier score which was optimized by GE. Importantly, the constructed features also benefited other performance metrics with an increase in area under ROC, balanced accuracy, and F1-Score. Interestingly, for two datasets, Diabetes and Crabs, all the metrics improved considerably. For the two datasets with no clear improvement, Irish and Image, the performance of NB with constructed features was similar to the base model. These results suggest that our framework was able to evolve feature representations which improved overall classification efficacy. Incorporating information about the interactions between features seems to increase the discriminative validity of the evolved representation, in comparison to base features. Furthermore, the versatile set of classification tasks at hand suggests that our framework is generally applicable in diverse research domains.

TABLE III. *AVERAGE PERFORMANCE ON UCI DATASETS*

Dataset	Model	ROC	Acc	F1	Brier	Runs
Breast	Base	0.992	0.955	0.969	0.034	4/5
	GE	**0.997**	**0.973**	**0.980**	**0.026**	
Pima	Base	0.800	0.693	0.798	0.215	4/5
	GE	**0.805**	**0.709**	**0.804**	**0.193**	
Heart	Base	0.884	0.787	0.798	0.162	5/5
	GE	0.873	**0.822**	**0.847**	**0.153**	
Irish	Base	0.640	0.600	0.542	0.241	2/5
	GE	0.596	0.570	0.504	0.247	
Image	Base	0.965	0.917	0.915	0.065	2/5
	GE	**0.977**	0.915	0.914	0.066	
Tennis	Base	0.497	0.451	0.526	0.425	4/5
	GE	0.467	**0.489**	**0.546**	**0.369**	
Diabet.	Base	0.960	0.925	0.881	0.073	4/5
	GE	**1**	**1**	**1**	**0.007**	
Crabs	Base	0.646	0.590	0.591	0.295	5/5
	GE	**0.982**	**0.900**	**0.897**	**0.066**	

B. Alzheimer's diagnosis

Results on the clinical problem of AD diagnosis confirmed our observations on the benchmark datasets. In Table IV the results from the application of our framework to the classification of AD vs HC are presented. Remarkably, the representation evolved by our framework exceeded the baseline performance on all four metrics in all five runs. It is evident that the constructed features greatly benefited classification. Therefore, it is plausible to suggest, that the consideration of feature interaction provides additional discriminative information.

TABLE IV. DIAGNOSIS RESULTS ON AD DATASET

Dataset	Model	ROC	Acc	F1	Brier	Runs
AD	Base	0.850	0.782	0.818	0.214	5/5
	GE	**0.913**	**0.815**	**0.859**	**0.178**	

In addition to our first goal of improving classification accuracy, our second goal was to maintain the explainability/human-interpretability of the evolved representations. To establish the interpretability of the GE constructed features, we investigated their biological relevance based on known results from the literature. Exemplary, two of the CF are discussed here. The first feature was constructed during the fourth run of our approach and represents a combination of three base features $CF_1 = x_{29} - x_{116} * x_{136}$. Although this CF integrates information about complex interactions into the model, the relation between the constituent base features is still understandable. The underlying brain regions were the left temporal pole (TmP, x_{29}), the ventromedial putamen (vmPu, x_{116}), and the right lateral prefrontal thalamus (lpThal, x_{136}). The location of these regions is depicted in Fig. 2. CF_1 suggests an interaction between TmP, lpThal and the vmPu and all are known to be affected in AD [31], [32], [33]. A second constructed feature, CF_2, represented a combination of additional three regions, such that $CF_2 = X_{135} + X_{104} * X_{128}$ (Fig.3). In this case, the underlying regions were the left lateral prefrontal thalamus (lpThal, X_{135}), the right lateral amygdala (lAmyg, X_{104}) and the right rostral temporal thalamus (rTThal, X_{128}). Apart from lpThal and rTThal for which we have already shown involvement in AD, the lAmyg is another region that is highly affected during the disease [34], [35]. The association of regional atrophy or co-atrophy in different brain regions may hint at the underlying biological mechanisms playing a role in development and course of AD. On average, the five runs on the ADNI data produced 130 features. A set of selected expressions from the runs are shown in Table V. Put in a clinical context, our approach is well suited to identify disease-relevant patterns. It is evident, that our proposed method is not only interpretable, but the basis on which classification is performed can be easily explained by analyzing the CFs.

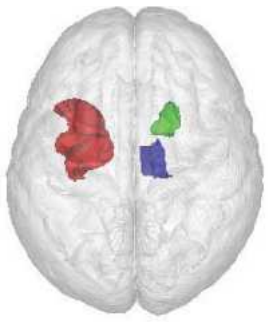


Fig. 2 Superior view of the brain. Depicted are TmP (red), vmPu (green) and lpThal (blue).

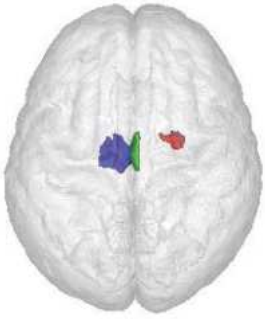


Fig. 3 Superior view of the brain. Depicted are lpThal (blue), lAmyg (red) and rTThal (green).

TABLE V. SELECTED EXPRESSIONS FROM 5 GE RUNS

Expression
$CF_1 = X_{29} - X_{116} * X_{136}$
$CF_2 = X_{135} + X_{104} * X_{128}$
$CF_3 = X_{38} - X_{59} / X_{122}$
$CF_4 = X_{76} - X_{132} * X_{83}$
$CF_5 = X_{136} - X_{53} * X_{74}$
$CF_6 = X_{101} * X_{119} / X_{66}$
$CF_7 = X_{48} / X_{23} * X_{56} / X_{101}$
$CF_8 = X_{110} + X_{23} - X_{158}$
$CF_9 = X_{51} - X_{46} * X_{65}$
$CF_{10} = X_{63} / X_{45} * X_{56}$

C. *Alzheimer's prognosis*

Since AD is marked by a continuous loss of neurons, early detection will play a vital role in future therapeutic methods. To this end, we tested if the diagnostic models using the features constructed for AD vs HC classification in the previous section would also be suitable for prognosis—i.e. to detect if MCI patients will later on convert to AD. If confirmed, it will indicate the generalizability of the constructed features and speak for their biological meaningfulness.

MCI is a neurological disorder that involves cognitive decline beyond what is expected for a person's age. It is generally seen as a prodromal stage of dementia, especially of AD [36]. Since not all MCI patients transition to dementia, it is a constant endeavor to differentiate between subjects on the verge of transitioning to AD (so-called converters), from stable MCI patients (non-converters). As the constructed features of our approach were able to pick up disease-relevant patterns in AD, we hypothesized that the same patterns could be useful to differentiate MCI-converters (MCIc) from stable MCI (MCIs) patients. Therefore, we extracted the same 173 features from 138 MCIc and 138 MCIs subjects' sMRI images from ADNI. The classification was performed first with the base model (trained on AD vs HC diagnostic data) and then using each of the five grammar models separately (again, trained on AD vs HC). Before applying the GE derived NB models, base features were transformed to match the constructed features of the respective model. In Table VI, the results of base and GE models are shown. The results of both models are comparable to those found in recent literature, although on the lower end of performance [37][38][39]. Nevertheless, our GE models could improve classification performance in comparison to the base model on all four metrics. Since GE is not limited to naïve Bayes classifiers, but well compatible with more sophisticated learning algorithms, future applications might yield even better results.

TABLE VI. PROGNOSIS RESULTS ON MCI DATASET

Dataset	Model	ROC	Acc	F1	Brier	Runs
ADNI	Base	0.717	0.680	0.699	0.316	5/5
	Grammar	**0.744**	**0.688**	**0.707**	**0.305**	

VI. CONCLUSION

We presented a simple GE based framework to evolve complex yet interpretable feature representations and showed its effectiveness on several benchmark datasets. We then tested the framework on two clinically relevant problems, diagnosis and prognosis of AD. In both cases, GE constructed features provided improved classification over base features. Moreover, the constructed features were interpretable. Our framework could prove useful in translational applications like the ones showcased here by providing both accuracy and interpretability.

Our framework is not without limitations. Firstly, we only considered the NB classifier primarily for its desirable property of low variance. However, other algorithms could provide higher accuracy if their variance can be properly controlled and should be tested. Secondly, we did not take special precautions to avoid overfitting in the optimization process itself [21], though our framework validates optimized models on hold-out data to avoid optimistic estimates. Heuristics such as early stopping could be investigated possibly further improving performance. Additionally, there can be multiple evolved solutions that perform equally well, which was the case for our results. In such cases, it is important to choose the most interpretable representation, which can be challenging.

Taken together, our simple framework can be useful for generating complex yet interpretable feature representations that can help improve both accuracy and interpretability.

ACKNOWLEDGMENT

This study was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 785907 (HBP SGA2) and Grant Agreement No. 7202070 (HBP SGA1). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.;

Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] V. H. Buch, I. Ahmed, and M. Maruthappu, "Artificial intelligence in medicine: current trends and future possibilities," *Br J Gen Pract*, vol. 68, no. 668, pp. 143–144, Mar. 2018, doi: 10.3399/bjgp18X695213.

[2] F. Jiang *et al.*, "Artificial intelligence in healthcare: past, present and future," *Stroke Vasc Neurol*, vol. 2, no. 4, pp. 230–243, Dec. 2017, doi: 10.1136/svn-2017-000101.

[3] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608 [cs, stat]*, Mar. 2017, Accessed: Dec. 10, 2019. [Online]. Available: http://arxiv.org/abs/1702.08608.

[4] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, May 2018, pp. 0210–0215, doi: 10.23919/MIPRO.2018.8400040.

[5] "EU guidelines on ethics in artificial intelligence: Context and implementation," Sep. 2019, [Online]. Available: http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)640163.

[6] B. Heinrichs and S. B. Eickhoff, "Your evidence? Machine learning algorithms for medical diagnosis and prediction," *Hum Brain Mapp*, p. hbm.24886, Dec. 2019, doi: 10.1002/hbm.24886.

[7] Anirban Mukhopadhyay, David Kügler, Andreas Bucher, Dieter Fellner, Thomas Vogl, "Putting Trust First in the Translation of AI for Healthcare," Jan. 22, 2019.

[8] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, p. 78, Oct. 2012, doi: 10.1145/2347736.2347755.

[9] M. O'Neill and C. Ryan, "Grammatical evolution," *IEEE Trans. Evol. Computat.*, vol. 5, no. 4, pp. 349–358, Aug. 2001, doi: 10.1109/4235.942529.

[10] A. M. D. Silva, F. Noorian, R. I. A. Davis, and P. H. W. Leong, "A Hybrid Feature Selection and Generation Algorithm for Electricity Load Prediction Using Grammatical Evolution," in *2013 12th International Conference on Machine Learning and Applications*, Miami, FL, USA, Dec. 2013, pp. 211–217, doi: 10.1109/ICMLA.2013.125.

[11] D. Gavrilis, I. G. Tsoulos, and E. Dermatas, "Selecting and constructing features using grammatical evolution," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1358–1365, Jul. 2008, doi: 10.1016/j.patrec.2008.02.007.

[12] O. Smart, I. G. Tsoulos, D. Gavrilis, and G. Georgoulas, "Grammatical Evolution for Features of Epileptic Oscillations in Clinical Intracranial Electroencephalograms," *Expert Syst Appl*, vol. 38, no. 8, pp. 9991–9999, Aug. 2011, doi: 10.1016/j.eswa.2011.02.009.

[13] A. A. Motsinger, D. M. Reif, S. M. Dudek, and M. D. Ritchie, "Understanding the Evolutionary Process of Grammatical Evolution Neural Networks for Feature Selection in Genetic Epidemiology," in *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, Toronto, Ont., Sep. 2006, pp. 1–8, doi: 10.1109/CIBCB.2006.330945.

[14] G. Georgoulas, D. Gavrilis, I. G. Tsoulos, C. Stylios, J. Bernardes, and P. P. Groumpos, "Novel approach for fetal heart rate classification introducing grammatical evolution," *Biomedical Signal Processing and Control*, vol. 2, no. 2, pp. 69–79, Apr. 2007, doi: 10.1016/j.bspc.2007.05.003.

[15] C. Patterson, "World Alzheimer Report 2018 - The state of the art of dementia research."

[16] S. Kloppel *et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, Feb. 2008, doi: 10.1093/brain/awm319.

[17] M. Liu, D. Zhang, D. Shen, and Alzheimer's Disease Neuroimaging Initiative, "Ensemble sparse classification of Alzheimer's disease," *Neuroimage*, vol. 60, no. 2, pp. 1106–1116, Apr. 2012, doi: 10.1016/j.neuroimage.2012.01.055.

[18] A. V. Lebedev *et al.*, "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," *Neuroimage Clin*, vol. 6, pp. 115–125, 2014, doi: 10.1016/j.nicl.2014.08.023.

[19] Trevor Hastie Robert Tibshirani Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, Second Edition. 2009.

[20] E. M. Dos Santos, R. Sabourin, and P. Maupin, "Overfitting cautious selection of classifier ensembles with genetic algorithms," *Information Fusion*, vol. 10, no. 2, pp. 150–162, Apr. 2009, doi: 10.1016/j.inffus.2008.11.003.

[21] J. Loughrey and P. Cunningham, "Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets," in *Research and Development in Intelligent Systems XXI*, M. Bramer, F. Coenen, and T. Allen, Eds. London: Springer London, 2005, pp. 33–43.

[22] Gavin C. Cawley, Nicola L.C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *JMLR*.

[23] Joaquín Abellán and Javier Castellano, "Improving the Naive Bayes Classifier via a Quick Variable Selection Method Using Maximum of Entropy," *Entropy*, vol. 19, no. 6, p. 247, May 2017, doi: 10.3390/e19060247.

[24] F. Noorian, A. M. de Silva, and P. H. W. Leong, "**gramEvol** : Grammatical Evolution in *R*," *J. Stat. Soft.*, vol. 71, no. 1, 2016, doi: 10.18637/jss.v071.i01.

[25] John H. Holland, "Genetic Algorithms," *Scientific American*, vol. 267, pp. 66–73, Jul. 1992.

[26] Glenn W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Wea. Rev.*, vol. 78, pp. 1–3, 1950.

[27] R. C. Petersen *et al.*, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010, doi: 10.1212/WNL.0b013e3181cb3e25.

[28] A. Schaefer *et al.*, "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI," *Cerebral Cortex*, vol. 28, no. 9, pp. 3095–3114, Sep. 2018, doi: 10.1093/cercor/bhx179.

[29] L. Fan *et al.*, "The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture," *Cereb. Cortex*, vol. 26, no. 8, pp. 3508–3526, Aug. 2016, doi: 10.1093/cercor/bhw157.

[30] R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. T. Yeo, "The organization of the human cerebellum estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, no. 5, pp. 2322–2345, Nov. 2011, doi: 10.1152/jn.00339.2011.

[31] L. W. de Jong *et al.*, "Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study," *Brain*, vol. 131, no. 12, pp. 3277–3285, Dec. 2008, doi: 10.1093/brain/awn278.

[32] M. Zarei *et al.*, "Combining shape and connectivity analysis: An MRI study of thalamic degeneration in Alzheimer's disease," *NeuroImage*, vol. 49, no. 1, pp. 1–8, Jan. 2010, doi: 10.1016/j.neuroimage.2009.09.001.

[33] J. Hänggi, J. Streffer, L. Jäncke, and C. Hock, "Volumes of Lateral Temporal and Parietal Structures Distinguish Between Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease," *JAD*, vol. 26, no. 4, pp. 719–734, Oct. 2011, doi: 10.3233/JAD-2011-101260.

[34] C.-A. Cuénod, "Amygdala Atrophy in Alzheimer's Disease: An In Vivo Magnetic Resonance Imaging Study," *Arch Neurol*, vol. 50, no. 9, p. 941, Sep. 1993, doi: 10.1001/archneur.1993.00540090046009.

[35] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, and B. C. Dickerson, "Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity," *Psychiatry Research: Neuroimaging*, vol. 194, no. 1, pp. 7–13, Oct. 2011, doi: 10.1016/j.pscychresns.2011.06.014.

[36] S. Gauthier *et al.*, "Mild cognitive impairment," *The Lancet*, vol. 367, no. 9518, pp. 1262–1270, Apr. 2006, doi: 10.1016/S0140-6736(06)68542-5.

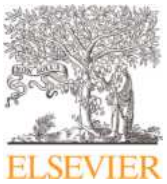
[37] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *NeuroImage*, vol. 104, pp. 398–412, Jan. 2015, doi: 10.1016/j.neuroimage.2014.10.002.

[38] H.-I. Suk and D. Shen, "Deep Learning-Based Feature Representation for AD/MCI Classification," in *Advanced Information Systems Engineering*, vol. 7908, C. Salinesi, M. C. Norrie, and Ó. Pastor, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 583–590.

[39] S. Basaia *et al.*, "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," *NeuroImage: Clinical*, vol. 21, p. 101645, 2019, doi: 10.1016/j.nicl.2018.101645.

3 Naturalistic viewing increases individual identifiability based on connectivity within functional brain networks, Jean-Philippe Kröll, Patrick Friedrich, Xuan Li, Kaustubh R. Patil, Lisa Mochalski, Laura Waite, Xing Qian, Michael WL Chee, Juan Helen Zhou, Simon Eickhoff, Susanne Weis, *NeuroImage, Vol*(273), (2023)

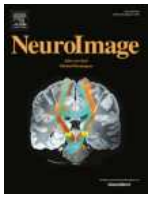
NeuroImage 273 (2023) 120083



Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



Naturalistic viewing increases individual identifiability based on connectivity within functional brain networks

Jean-Philippe Kröll[a,b,*], Patrick Friedrich[a,b], Xuan Li[a,b], Kaustubh R. Patil[a,b], Lisa Mochalski[a,b], Laura Waite[a], Xing Qian[c], Michael WL Chee[c,e], Juan Helen Zhou[c,d,e], Simon Eickhoff[a,b], Susanne Weis[a,b]

[a] Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Jülich, Jülich 52428, Germany
[b] Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany
[c] Yong Loo Lin School of Medicine, Centre for Sleep and Cognition and Centre for Translational MR Research, National University of Singapore, Singapore
[d] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[e] Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore

A B S T R A C T

Naturalistic viewing (NV) is currently considered a promising paradigm for studying individual differences in functional brain organization. While whole brain functional connectivity (FC) under NV has been relatively well characterized, so far little work has been done on a network level.

Here, we extend current knowledge by characterizing the influence of NV on FC in fourteen meta-analytically derived brain networks considering three different movie stimuli in comparison to resting-state (RS). We show that NV increases identifiability of individuals over RS based on functional connectivity in certain, but not all networks. Furthermore, movie stimuli including a narrative appear more distinct from RS. In addition, we assess individual variability in network FC by comparing within- and between-subject similarity during NV and RS. We show that NV can evoke individually distinct NFC patterns by increasing inter-subject variability while retaining within-subject similarity. Crucially, our results highlight that this effect is not observable across all networks, but rather dependent on the network-stimulus combination. Our results confirm that NV can improve the detection of individual differences over RS and underline the importance of selecting the appropriate combination of movie and cognitive network for the research question at hand.

1. Introduction

Understanding functional brain organization is a major goal of human neuroscience. Typically, researchers have focused on commonalities between individuals and used group-averages to reveal the shared neural underpinnings of certain brain functions. In recent years, the interest in individual functional brain architecture has grown. At the same time, neuroimaging has shifted from mapping brain functions towards investigating interactions within distributed brain networks by considering functional brain connectivity. Specifically, functional connectivity studies yielded insight into the foundation of individual brain organization (Biswal et al., 1995; Greicius et al., 2003; Fox et al., 2006; Damoiseaux et al., 2006). However, it is yet unclear which paradigms are best suited to study individual differences.

Most research on FC has been done on connectivity patterns occurring during resting state (RS), where participants lie in the scanner without any particular task or any external stimulation (Damoiseaux et al., 2006; Amft et al., 2015; Langner and Eickhoff, 2013; Binder et al., 2009; Buhle et al., 2014; Shehzad et al., 2009; Schaefer et al., 2018). In contrast to task-based studies, RS is thought to reveal the intrinsic brain organization (Smith et al., 2009). In addition, the ease of implementation of RS data allows for the relatively quick acquisition of large healthy and clinical samples due to low demands on participants. Although the RS paradigm has provided a variety of insights into the organization of the human brain, it also comes with limitations: In the absence of a task, RS is likely influenced by spontaneous thoughts of the participant (Christoff et al., 2004; Gonzalez-Castillo et al., 2021). Furthermore, experimental decisions such as instructing participants to keep their eyes open or closed can affect the measurement (Patriat et al., 2013). Finally, various studies have shown that individual FC during RS is heavily influenced by state effects (Geerligs et al., 2015).

To address these limitations, naturalistic viewing (NV) has been suggested as a promising tool for the study of individual differences (Finn et al., 2017; Finn et al., 2020). During NV, participants are instructed to watch a movie clip without any additional task. Therefore, NV reduces the variability induced by spontaneous thought content of the subject, because all participants are presented with the same stimulus (Hasson et al., 2004). By more closely mimicking conditions under which the brain naturally operates, NV promises to capture more ecologically valid neuronal responses. Despite NV increasing the similarity of FC across participants, individual differences still persist. Using “fingerprinting” (Finn et al., 2015) or identifiability as a proxy for individual

* Corresponding author at: Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Jülich, Jülich 52428, Germany.
E-mail address: j.kroell@fz-juelich.de (J.-P. Kröll).

https://doi.org/10.1016/j.neuroimage.2023.120083.
Received 26 October 2022; Received in revised form 7 March 2023; Accepted 31 March 2023
Available online 2 April 2023.

differences, Vanderwal et al., (2007) demonstrated that NV shows better identification accuracy than RS (Vanderwal et al., 2017). Furthermore, Finn et al. (2020) showed that the implementation of NV data outperforms RS in predicting trait-like phenotypes, thus suggesting that individual variability might be enhanced during NV (Finn and Bandettini, 2020). Different attempts have been made to explain why NV might enhance FC variability. For instance, Geerligs et al. (2015) argued that the differences in interpretation of a given movie content might promote individual FC variability (Geerligs et al., 2015). Van de Meer and colleagues (der et al., 2020) suggested that NV might impose richer brain state dynamics and therefore more distinct connectivity profiles, which in turn might better reflect phenotypes of interest than brain states during RS. Naturalistic Viewing paradigms provide further advantages over conventional RS: By increasing participant engagement, NV reduces fatigue and head movement during the measurement (Finn and Bandettini, 2020; Vanderwal et al., 2019). In addition, movie-watching can increase scanner tolerability for cohorts which might either struggle with staying still (e.g. ADHD patients) or completing demanding tasks (subjects with cognitive impairments) (Eickhoff et al., 2020).

Current literature evinces the potential for naturalistic viewing as a useful paradigm for studying individual brain architecture. So far, most studies primarily focused on whole-brain connectivity reflecting a holistic view on brain functions. However, brain architecture is commonly seen as segregated into modular clusters of spatially distinct areas constituting functional networks (Sporns and Betzel, 2016). These networks represent specific cognitive domains, such as memory (Spreng et al., 2009), social cognition (Bzdok et al., 2012) and executive function (Rottschy et al., 2012). Therefore, investigating networks functional connectivity (NFC) increases the interpretability of findings over whole-brain connectivity. Furthermore, connectivity in different networks likely yields distinct patterns of variance in reaction to NV stimuli. For example, a functional network related to the processing of emotions should react differently to a movie scene with strong emotional content, as compared to the motor network.

The most commonly used method to define functional networks is to estimate them from FC under resting-state (Damoiseaux et al., 2006; Schaefer et al., 2018; Thomas Yeo et al., 2011). RS-networks have shown good reproducibility and seem to generally converge well with studies on task-evoked networks (Smith et al., 2009; Mennes et al., 2010; Dosenbach et al., 2007). However, there are several other methods for defining functional networks (Schaefer et al., 2018; Smith et al., 2009; Power et al., 2011), one of which are meta-analytically defined networks (Eickhoff et al., 2012). The latter have the advantage of representing the most likely core nodes involved in a given cognitive function, because they incorporate convergent information from a multitude of studies (Eickhoff et al., 2020). Thus, studying NFC in meta-analytical networks might grant robust insights into the effects of naturalistic viewing on individual variability, which has not been studied yet.

The present study aims to investigate the influence of NV on individual variability in NFC by use of three different movie stimuli and RS. There is a plethora of NV stimuli available. Depending on the research question at hand, studies have suggested to use stimuli that are disease-specific (e.g. a movie evoking suspicion to study paranoia) (Eickhoff et al., 2020; Finn et al., 2018), emotionally or socially engaging (Finn and Bandettini, 2020; Saarimäki, 2021; Mishra et al., 2022; Schaefer et al., 2010) or as neutral as possible (Vanderwal et al., 2015). Previous studies on individual variability under NV employed stimuli that the researchers deemed to be the most engaging, thus resorting to movies with high social and emotional content (Finn and Bandettini, 2020; Saarimäki, 2021; Mishra et al., 2022; Schaefer et al., 2010). We employ stimuli with different levels of social content, ranging from the neutral movie *Inscapes*, over the silent movie *The Circus,* to the most social movie *Indiana Jones and the Temple of Doom.* Understanding how different levels of social and emotional content influence individual variability on a network level might aid researchers in choosing adequate stimuli for future studies.

We compare several measures of individual variability (e.g. identifiability and inter- and intra-subject variability) between the three different movie stimuli and RS across three scanning sessions on the basis of various meta-analytical networks covering affective (Amft et al., 2015; Buhle et al., 2014; Liu et al., 2011; Sabatinelli et al., 2011), social (Amft et al., 2015; Bzdok et al., 2012; Caspers et al., 2010), executive (Langner and Eickhoff, 2013; Rottschy et al., 2012; Camilleri et al., 2018; Cieslik et al., 2015), memory (Binder et al., 2009; Spreng et al., 2009) and motor (Witt et al., 2008) functions. Furthermore, we validate our results in RS-derived networks by Thomas Yeo et al. (2011), and on a whole-brain atlas by Shen et al. (2013). As a first step, we examined the similarity of connectivity profiles evoked by different movies and RS. Secondly, we assessed the identifiability of subjects based on NFC-patterns evoked by NV or RS. Subsequently, we investigated to what extent identifiability is affected by network size. Finally, we compared the effect of different movies and RS on inter- and intra-subject variability.

2. Material and methods

2.1. Participants

36 healthy right-handed and ambidextrous adults were scanned at the centre for Translational MR Research, National University of Singapore. Two subjects were excluded for having incomplete sessions, leaving a final cohort at 34 (19 females, mean age 27 + / - 2.7 years). Exclusion criteria were neurological or psychiatric diagnoses, significant visual or hearing impairment, alcohol or caffeine consumption 6 h prior to the scan and self-reporting of bad sleep the night before the scan days. All participants underwent three identical testing sessions within a one-week interval. Subjects gave written, informed consent and were compensated for their participation. The study was approved by the institutional review board of the National University of Singapore.

2.2. Data acquisition

The data was acquired on a Siemens Magnetom PrismaFit 3-Tesla with a 20-Channel head coil. Structural images were collected using an MP-RAGE sequence (TR=2300 ms, TE =2,28 ms, TI=900 ms, flip-angle=8°) and 1 mm voxel size. All RS and NV runs used the same echo planar imaging sequence (TR=719 ms, TE=30 ms, flip-angle=52°, slices=44, FOV=225×225 mm^2) resulting in 2.96×2.96×3 mm voxel size. Data were retrieved from collaborators at the National University of Singapore, and structured in the form of a DataLad dataset, a research data management solution providing data versioning, data transport, and provenance capture Halchenko et al. (2021). Each of the three testing sessions per participant, which were conducted within a seven day period, comprised three NV runs and two RS scans. The order of scans was identical on all three days, starting with a structural scan, followed by 5 functional scans in the order of RS 1, *Inscapes, Circus, Indiana Jones* and RS 2, with each functional scan lasting for 10 min. All movies had been cut to the same length. For RS scans, participants were asked to lay as still as possible and think of nothing in particular, while keeping their eyes open. Instructions for the NV scans were to watch the movies while staying as still as possible. For all scans, participants were asked to not fall asleep during the measurement. The movie clips were presented via a mirror that was mounted on the head coil and the sound was played through headphones. *Inscapes* is a non-verbal, non-social series of animated abstract shapes created by Vanderwal et al. which was looped to match the 10 min duration (original length 7 min) (Vanderwal et al., 2015). *The Circus* (United Artists Digital Studios, 1928, directed by Charlie Chaplin) is a silent black-and-white film which depicts the protagonist being chased by the police and unintentionally causing comic situations during his escape. *Indiana Jones and the Temple of Doom* (Paramount Pictures, 1984, directed by Steven Spielberg) shows the opening scene of the movie during which the protagonist has to fight off several hitmen who are trying to kill him. Foam

wedges were fitted around each subject's head for comfort and to decrease movement. For all subsequent analyses, only the first RS scan (RS1) was used.

2.3. Data preprocessing

Preprocessing of MRI data was performed using fMRIPrep, version 20.1.1 (Esteban et al., 2019). In brief, the T1-weighted volumes were corrected for intensity non-uniformity and skull-stripped. The extracted brain images were then transformed into Montreal Neurological Institute (MNI) space and motion corrected using Advanced Normalization Tools (ANTS) (Avants et al., 2009). The functional data was motion-corrected with MCflirt (Jenkinson et al., 2002) and subsequently co-registered to the native T1-weighted image using boundary based registration with six degrees of freedom from Freesurfer (Greve and Fischl, 2009). Subsequently, ICA-AROMA (Pruim et al., 2015) was used on the MNI-aligned BOLD images to remove motion artifacts and applied an isotropic Gaussian kernel of 6 mm FWHM (full-width half-maximum) for spatial smoothing. Global signals were extracted within the CSF, the WM, and the whole-brain masks and regressed from the preprocessed fMRI data for each subject.

2.4. Network functional connectivity

For each subject, NFC matrices were constructed for each of the 14 meta-analytical networks, comprising nine to 23 nodes (a detailed description of the networks can be found in the supplements). Isotropic 5 mm spheres were created around the local maxima of each meta-analytical network node and the mean time series were subsequently extracted. Only gray matter voxels were included. In addition, NFC matrices were constructed for the seven RS derived networks created by Thomas Yeo et al. (2011), comprising the Default, Control, Dorsal Attention, Salience, Visual, Somatomotor and Limbic networks, and the whole-brain atlas created by Shen et al. (2013). Pearson's correlation coefficient (PCC) between all node pairs was calculated to generate a n-times-n connectivity matrix per subject and condition, where n denotes the number of nodes of the respective network.

2.5. Representational dissimilarity matrix (RDM) analysis

To investigate how patterns of inter-individual differences in NFC vary across conditions (RS and three different NV conditions), we applied a RDM analysis. The present analysis closely followed the methods described by Kriegeskorte (2008). The procedure can be summarized in three steps. First, the correlation between the FC patterns of every possible pair of subjects is calculated for each condition and network. Second, to generate a measure of dissimilarity, the correlation distance (1-r) is computed. Third, the dissimilarity values for all subject pairs are assembled into an RDM (as a subjects * subjects size matrix) that serves as the signature of the given condition.

To visually compare RDMs, we employed Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), a technique for dimensionality reduction and visualization. Instead of preserving large-scale structures, UMAP seeks to preserve local neighborhood distances. To this end, a pre-set number of nearest neighbors (NN) are specified and the distances to these neighbors is represented as a weighted graph, with the NN being assigned with higher weights. UMAP then finds a low-dimensional representation of the data that best preserves these neighborhoods. The NN parameter controls whether UMAP focuses on the local or global structure of the data. Large values force UMAP to consider a larger number of neighbors and therefore focus on the broader structure of the data. In contrast, low values of NN force UMAP to focus on the local structure of the data. We here applied UMAP to the previously described RDMs. To account for the small number of data points (fifteen RDMs per network) the NN parameter was set to two. Considering more than 4 NN led to a more global clustering of RDMs that partly obscured differences between conditions. By grouping closely related RDMs together, UMAP allowed us to visualize which conditions evoked similar responses. Of note, distance metrics in UMAP are non-linear and not necessarily the same for each dimensionality reduction. Therefore, the results are suited to compare the similarity of condition evoked responses within, but not across networks. An analysis of the RDMs on a whole-brain level is reported in the supplementary material (Fig. S1).

2.6. Assessment of identifiability

Assessment of identifiability was closely based on the methods described by previous papers (Finn et al., 2015; Vanderwal et al., 2017). The FC matrices belonging to the same session and condition were grouped, resulting in 12 databases (three sessions times the four conditions). For every combination of two databases, Pearson's correlation between the FC matrix of one subject from the first database and every other FC matrix from the second database was calculated. The two FC matrices with the highest correlation were considered to be from the same subject. Identification accuracy was defined as the frequency of correctly identified subjects divided by the total number of subjects. Afterwards, the accuracies were averaged across session pairs to quantify the identifiability per condition and network. An analysis of identifiability on a whole-brain level is reported in the supplementary material (Table S1).

2.7. Influence of network size

To ensure that the differences in identification accuracy between networks were not just reflections of network size, we systematically compared identifiability in artificially created networks, constituting up to 50 nodes. Artificial networks were created by randomly choosing coordinates from the MNI152 gray matter mask. Around each coordinate, an isotropic sphere was created, which was matched to the node size of the meta-analytical networks (5 mm). The mean Euclidean distance between nodes from the meta-analytically defined networks was calculated (14.62 mm) and set as the minimal distance between nodes for the artificial networks. Thereby, the randomly chosen nodes were prevented from overlapping whilst preserving some degree of spatial comparability between artificial and meta-analytically defined networks. This process was repeated 100 times for each network size, creating a new random configuration of nodes during each repetition. Subsequently, identification accuracies for all networks and the different conditions were calculated to evaluate (1) how network size influences identification accuracy, (2) how identifiability between the different conditions behaves in artificial networks and (3) how the meta-analytically defined networks compare to the artificial networks.

2.8. Within- and between subject correlation

Within-subject correlations were calculated as Pearson's correlation between the FC matrices of the same subject across session pairs (e.g. Ses-1 to Ses-2, Ses-1 to Ses-3) and then averaged. This process was performed for each of the four conditions (RS and the three movie stimuli) separately. For each network or whole-brain atlas, a one-way ANOVA was computed with condition (RS, *Inscapes, Circus, Indiana Jones*) as within-subject factor to evaluate the effect of condition on within-subject correlations within the specific networks. Subsequently, Bonferroni correction was applied to account for Type 1 error and Tukey's HSD test was performed to reveal which of the conditions significantly differed. The between-subject correlations were defined as the mean PCC between the FC matrix of one subject and every other subject's FC matrix from the same session and condition. For each network, a one-way ANOVA was computed with condition (RS, movie1, movie2, movie3) as between-subject factor to evaluate the effect of condition on between-subject correlations within the specific networks. Subsequently, Bonferroni correction was applied to account for Type 1 error and Tukey's

HSD test was performed to reveal which of the conditions significantly differed. It is important to note that the between-subject comparisons in this study are based on correlations between static NFC of subjects, in contrast to an Inter-subject Correlation (ISC) approach that correlates the fMRI time series of subjects and is often used to analyze NV (Halchenko et al., 2021). As such, our results should not be interpreted as a measure of synchrony across subjects, but rather as their similarity in FC. The analysis of within- and between-subject correlations on a whole-brain level can be found in the supplementary material (Fig. S2).

3. Results

3.1. Similarity of different movies and RS connectivity profiles in meta-analytic networks

We investigated the similarity of different conditions by embedding the respective RDMs into a low dimensional space (UMAP).The UMAP representation showed that RS was embedded separately from all NV conditions in AM, CogAC, VigAtt and WM, and separately from most NV conditions in MNS and Motor networks. In eMDN, EmoSF, ER, eSAD, Rew and ToM networks, RS shows overlaps with the movie *Inscapes*. On the other hand, the movies *Circus* and *Indiana Jones* tended to cluster together in (AM, CogAC, eMDN, Empathy, ER, eSAD, MNS, Motor, Rew). We did not observe any evidence for a systematic session-effect, as RDMs of the same session (session 3) were only embedded together in the motor network (Fig. 1).

3.2. Similarity of different movies and RS connectivity profiles in RS derived networks

The UMAP representation of the different conditions in RS derived networks showed that RS was embedded separately from all NV conditions in the Control network and separately from most NV conditions in Limbic, SomatoMotor and Visual networks. In all networks except for the Control network, RS shows overlap with the movie *Inscapes. Indiana Jones* and *Circus* overlap in all networks (Fig. 2).

3.3. Identification accuracies in meta-analytic networks

Identifiability of subjects was assessed based on NFC-patterns evoked by NV or RS. Overall, individual FC matrices could be matched across sessions with moderate to high accuracy with identification accuracies ranging from 52% to 100%. The motor network represented an exception with low identification accuracies across conditions (27.5%–30.4%). In eleven out of 14 networks, identifiability was highest in either the *Circus* or *Indiana Jones* NV conditions. Among the naturalistic stimuli, *Indiana Jones* led to the highest identification accuracies in eight of the networks (SM, CogAc, EmoSF, eMDN, ER, VigAtt, MNS, and eSAD). The top three highest accuracies were achieved using NV, with FC matrices using the *Indiana Jones* movie reaching the highest accuracy (98%) in the SM network. Generally, networks with more nodes tended to achieve higher accuracies.

3.4. Identification accuracies in RS derived networks

In addition, identifiability of subjects was assessed based on NFC in RS derived networks. Generally, individual FC matrices could be matched with moderate to high accuracy with accuracies ranging from 43% to 91%. The limbic network represented an exception with low identification accuracies across conditions (9.3%–14.22%). In the control, dorsal attention and visual networks, *Indiana Jones* led to the highest identification accuracy. In the default, salience and somatomotor networks, RS led to the highest identifiability. The highest accuracy was achieved by RS in the default network (91%). Overall, accuracies in the RS derived networks were lower than in the majority of meta-analytically derived networks.

3.5. Identification accuracies for different network sizes

To evaluate the effect of network size on identification accuracy, we computed identifiability in random networks with sizes between 3 and 50 nodes. We then compared these to the accuracies achieved in meta-analytic networks, as the meta-analytic networks showed higher accuracy then the RS derived networks. Identifiability in artificial networks showed how network size influences identification accuracy for all modalities (Fig. 2). A continuous increase of identification accuracy can be seen for all conditions up until a network size of 20 nodes, where the increase rate stabilizes. All networks, apart from the Motor network, achieved higher accuracies than the artificially created networks of the same size, regardless of condition. Furthermore, identification accuracies for the *Indiana Jones* movie exceeded those of the other three conditions, regardless of network size (Fig. 3).

3.6. Within- and between-subject correlations in meta-analytic networks

We calculated within-subject correlations, as a measure of how similar subjects are to themselves across sessions, and between-subject correlations, as a measure of similarity between subjects. The average within-subject correlations for RS and NV ranged between 0.5 and 0.8, with the exception of the Motor network (0.1–0.6), indicating a high level of similarity of connectivity patterns across sessions. For multiple networks, most prominently the MNS network, within-subject correlations strengthened from RS < *Inscapes* < *Circus* < *Indiana Jones*.

RS state differed from one or more movie conditions in various networks: RS showed significantly higher within-subjects correlations compared to *Indiana Jones* (AM) and *Circus* (AM). In contrast, some movies showed significantly higher within-subject correlations than RS in emoSF (*Indiana Jones*), and MNS (*Indiana Jones* and *Circus*).

In several networks certain movies differed from one another, with significantly higher correlations in *Indiana Jones* compared to *Circus* in emoSF; and higher correlations in *Indiana Jones* compared to *Inscapes* in Empathy and MNS networks. *Circus* never showed significantly higher correlations compared to any other movie in any network.

RS and the movie *Inscapes* exhibited similar correlations across networks. Overall, the movie *Indiana Jones* tended to stand out in that it was the only condition that showed significantly higher within-subject correlations than RS in several networks (EmoSF and MNS). On the contrary, the movie *Circus* often led to decreased within-subject correlations in comparison to the other conditions.

Between-subject correlations were generally lower than those previously observed on a whole-brain level, ranging from below 0.1 to 0.75. In several networks, the opposite pattern of what was observed for within-subject correlations can be seen, such that increasingly complex stimuli weaken between-subject correlations (AM, ER, eSAD and SM). In other networks, the three movies made connectivity across subjects more similar, increasing between-subject correlations in comparison with RS (CogAc, EmoSF, Rew and VigAtt).

Comparing within- and between-subject correlations, it is evident that increased within-subject correlations did not automatically lead to decreased between-subject correlations (and vice versa), such that a subject's scan can be highly individual (or reliable) and still share substantial overlap with others.

RS differed from one or more movie conditions in various networks: RS showed significantly higher between-subjects correlations compared to *Indiana Jones* (AM, eSAD, SM, ToM), *Inscapes* (ToM) and *Circus* (AM, Motor, SM, ToM). In contrast, other networks showed significantly higher between-subject correlations than RS for *Indiana Jones* (CogAC, eMDN, EmoSF, Rew, VigAtt,WM), *Inscapes* (CogAC, emoSF, Rew, VigAtt, WM,) and *Circus* (CogAC, EmoSF, MNS, Rew, VigAtt).

In several networks certain movies differed from one another, with significantly higher between-subject correlations of *Inscapes* compared to *Circus* in the AM, CogAC, EmoSF, eSAD, Motor, SM, and ToM; and higher correlations in *Inscapes* compared to *Indiana Jones* in AM, EmoSF,

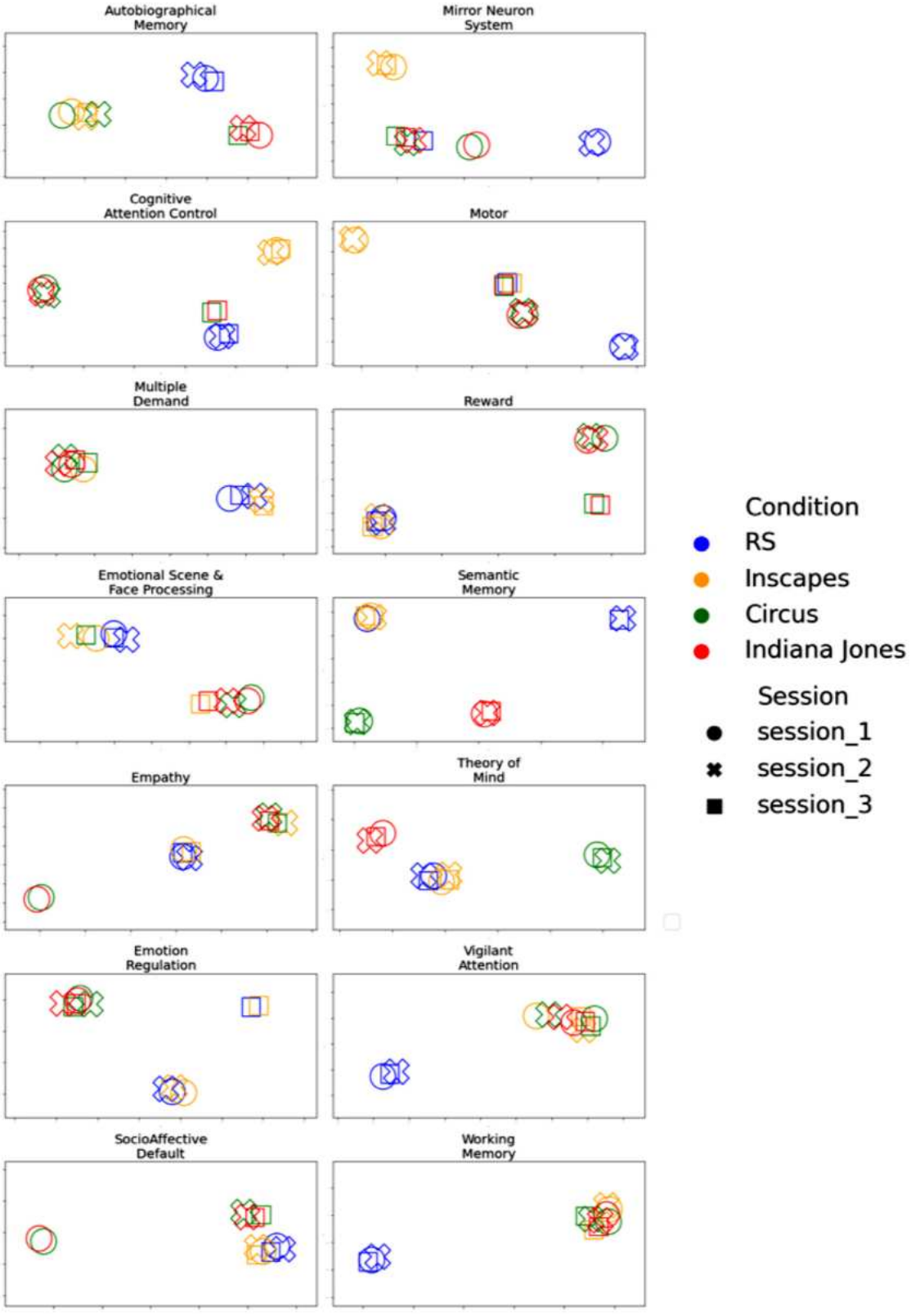


Fig. 1. UMAP representation of the RDMs of the different conditions in each meta-analytic network.

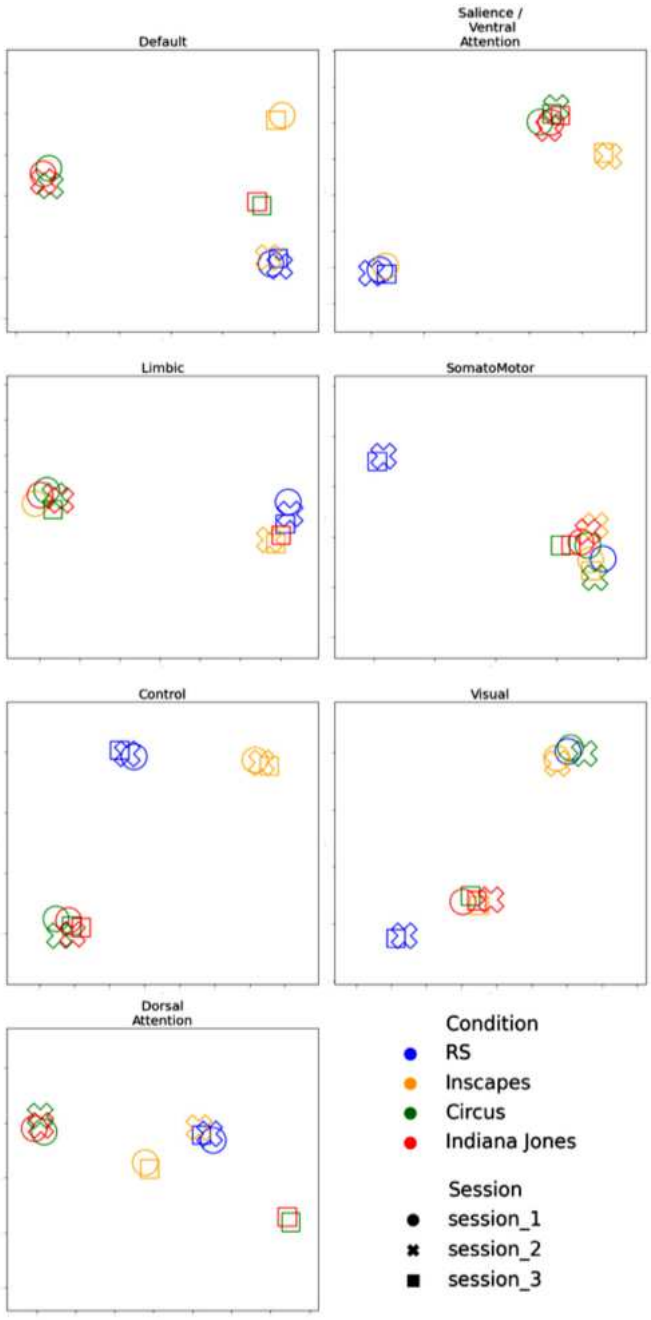


Fig. 2. UMAP representation of the RDMs of the different conditions in each RS derived network.

ER, eSAD and SM; and higher correlations in *Indiana Jones* compared to *Circus* in eDMN, Motor and ToM; and higher correlations in *Circus* compared to *Indiana Jones* in AM and SM networks (Figs. 4 and 5).

3.7. Within- and between-subject correlations in RS derived networks

We calculated within- and between-subject correlations for the RS derived networks. The average within-subject correlation for RS and NV ranged between 0.6 and 0.9, with the exception of the limbic network (0.1–0.8), indicating a high level of similarity of connectivity across sessions. The within-subject correlations in the RS derived networks were generally higher than the within-subject correlations in the meta-analytic networks. RS showed significantly higher within-subject correlations than *Circus* in the default network.The movie *Indiana Jones* showed significantly higher within-subject correlations than *Circus* in the Default network.

The average between-subject correlations ranged between 0.1 and 0.9 and were generally higher than the between-subject correlations in the meta-analytic networks. In five out of seven networks, at least one of the movie conditions led to higher between-subject correlations than for RS.

RS differed from one or movie conditions in various networks. RS showed significantly higher between-subject correlations compared to *Circus* (Cont) and *Indiana Jones* (Default). In contrast, other networks showed higher between-subject correlations than RS for *Inscapes* (DorsAtt), *Circus* (SalVentAtt, SomMot, Vis) and *Indiana Jones* (SalVentAtt, SomMot, Vis).

In several networks, certain movies differed from each other with significantly higher between-subject correlations for *Inscapes* than *Circus* in the Default and DorsAtt network; and higher correlations for *Inscapes* compared to *Indiana Jones* in the Default network; and higher correlations for *Circus* than *Inscapes* in the SalVentAtt, SomMot and Vis networks; and higher correlations for *Circus* than for *Indiana Jones* in the Vis network; and higher correlations for *Indiana Jones* than for *Inscapes* in the SalVentAtt and SomMot networks Figs. 6 and 7).

4. Discussion

In the current study we examined and compared the NFC evoked by different NV stimuli and RS with respect to similarity of connectivity profiles, individual identifiability, as well as within- and between-subject correlations. Our results showed that NV stimuli evoke connectivity profiles that are distinct from RS across meta-analytically defined and RS derived networks. NV stimuli, especially *Indiana Jones*, enhance the identifiability of individual subjects in the vast majority (10 of 14) of meta-analytic networks. Crucially, our results show that NFC analysis might reveal differences that are obscured on a whole brain basis. Lastly, our results emphasize that the similarity of individuals to themselves and to others is highly dependent on the combination of condition and network.

4.1. Comparison of connectivity profiles during NV and RS

In this study, we compared NFC evoked by three different NV stimuli and RS. A low-dimensional embedding of NFC similarity across subjects in meta-analytic networks showed that FC patterns during *Inscapes* are mostly similar to those during RS, while *Circus* and *Indiana Jones* exhibited distinct connectivity profiles across networks (Fig. 1). The relative similarity of connectivity patterns during Inscapes and RS has been reported before: For instance, based on Pearson's correlations between FC matrices, *Inscapes* was shown to be more similar to RS than to another movie condition (Vanderwal et al., 2017). These authors argued that due to the abstract nature of the movie, participants might not engage in temporally synchronized cognitive processes, which is similar to RS (Vanderwal et al., 2015). Furthermore, our embedding shows little similarity of NFC during *Inscapes* and either *Circus* or *Indiana Jones* in the majority of networks. This is in line with the previous argument, as both *Circus* and *Indiana Jones* contain a narrative that is likely to increase similarity across subjects, as has been shown for verbal narratives (e.g. emotional speeches (Nummenmaa et al., 2014; Schmälzle et al., 2015)). Accordingly, connectivity profiles during *Circus* and *Indiana Jones* overlap in the vast majority of networks. For the whole brain, similarity across conditions seemed more widespread and all conditions clustered together at least once (Fig. S1).

4.2. Identifiability

To assess the stability of individual patterns on the network level, we calculated the identifiability of NFC matrices across the three movies and RS (Table 1). Considering that NV has been shown to increase the

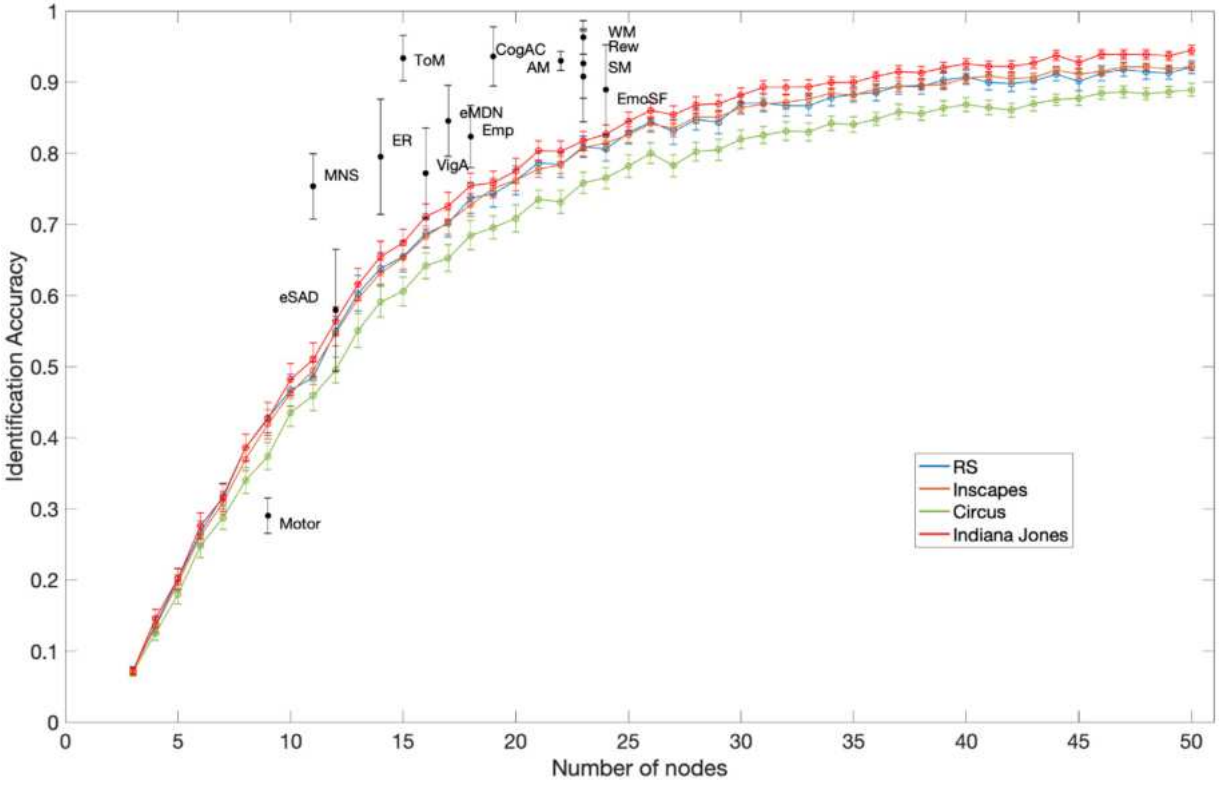


Fig. 3. Identification accuracies in artificial networks. The figure depicts the network size as the number of nodes (x-axis) against averaged identification accuracy (y-axis) for each of the four conditions (RS = blue; Inscapes = orange; Circus = green; Indiana Jones = red). Black dots denote the mean identification accuracy of meta-analytically defined networks, averaged across conditions and placed at their respective node count. (AM =Autobiographical Memory, CogAC = Cognitive Attention Control,eMDN=extended Multiple Demand Network, EmoSF= Emotional Scene and Face Processing, ER = Emotion Regulation, eSAD=Extended Social-affective Default, MNS = Mirror Neuron System, Rew = Reward, SM = Semantic Memory, ToM = Theory of Mind, VigAtt= Vigilant Attention, WM = Working memory,.

Table 1
Identification accuracies per network and modality, averaged across sessions. Networks are in order of highest average accuracy. The highest identification accuracy in each network is denoted in bold. (AM =Autobiographical Memory, CogAC = Cognitive Attention Control,eMDN=extended Multiple Demand Network, EmoSF= Emotional Scene and Face Processing, ER = Emotion Regulation, eSAD=Extended Social-affective Default, MNS = Mirror Neuron System, Rew = Reward, SM = Semantic Memory, ToM = Theory of Mind, VigAtt= Vigilant Attention, WM = Working memory, Shen = Shen atlas).

Network	RS	Inscapes	Circus	Indiana Jones	Node Number
Semantic Memory	95.1%	95.1%	97.1%	**98.0%**	23
Cognitive Attention Control	93.6%	90.2%	94.1%	**96.6%**	19
Theory of Mind	93.1%	90.7%	**95.1%**	94.6%	15
Autobiographical Memory	**94.1%**	92.2%	93.1%	92.6%	22
Working Memory	**96.1%**	93.6%	88.7%	92.2%	23
Reward	**96.1%**	90.7%	86.3%	90.2%	23
Emotional Scene & Face Perception	88.7%	85.8%	86.8%	**94.6%**	24
Multiple Demand Network	85.8%	85.8%	79.9%	**86.8%**	17
Empathy	**86.3%**	81.4%	79.9%	81.9%	18
Emotion Regulation	81.9%	80.9%	72.1%	**83.3%**	14
Vigilant Attention	80.4%	74.0%	73.5%	**80.9%**	16
Mirror Neuron System	**77.0%**	76.5%	71.1%	**77.0%**	11
Socio Affective Default	59.8%	52.9%	54.4%	**64.7%**	12
Motor	27.9%	30.4%	**30.4%**	27.5%	9

reliability of individual FC patterns (Geerligs et al., 2015; Hasson et al., 2010), we hypothesized that identifiability should be higher for movies as compared to RS. However, present results suggest that this is not the case for movies in general, but rather identification accuracy appears to highly depend on the specific movie as well as on the chosen network. Specifically, *Indiana jones* achieved the highest accuracy in 8 of 14 networks (SM, CogAC, EMOSF, eMDN, ER, VigAtt, MNS, eSAD), whereas *Inscapes* and *Circus* produced highest accuracies in two networks (*Inscapes*: Motor; *Circus*: ToM, Motor). RS, on the other hand, achieved the highest accuracies in 5 networks (AM, WM, ReW, Empathy, MNS). Notably, the connectivity profiles within the Motor network yielded low identification accuracies in comparison with the other networks across all stimuli. Lower-level cognitive structures such as the motor network show low variance between participants (Croxson et al., 2018). Further-

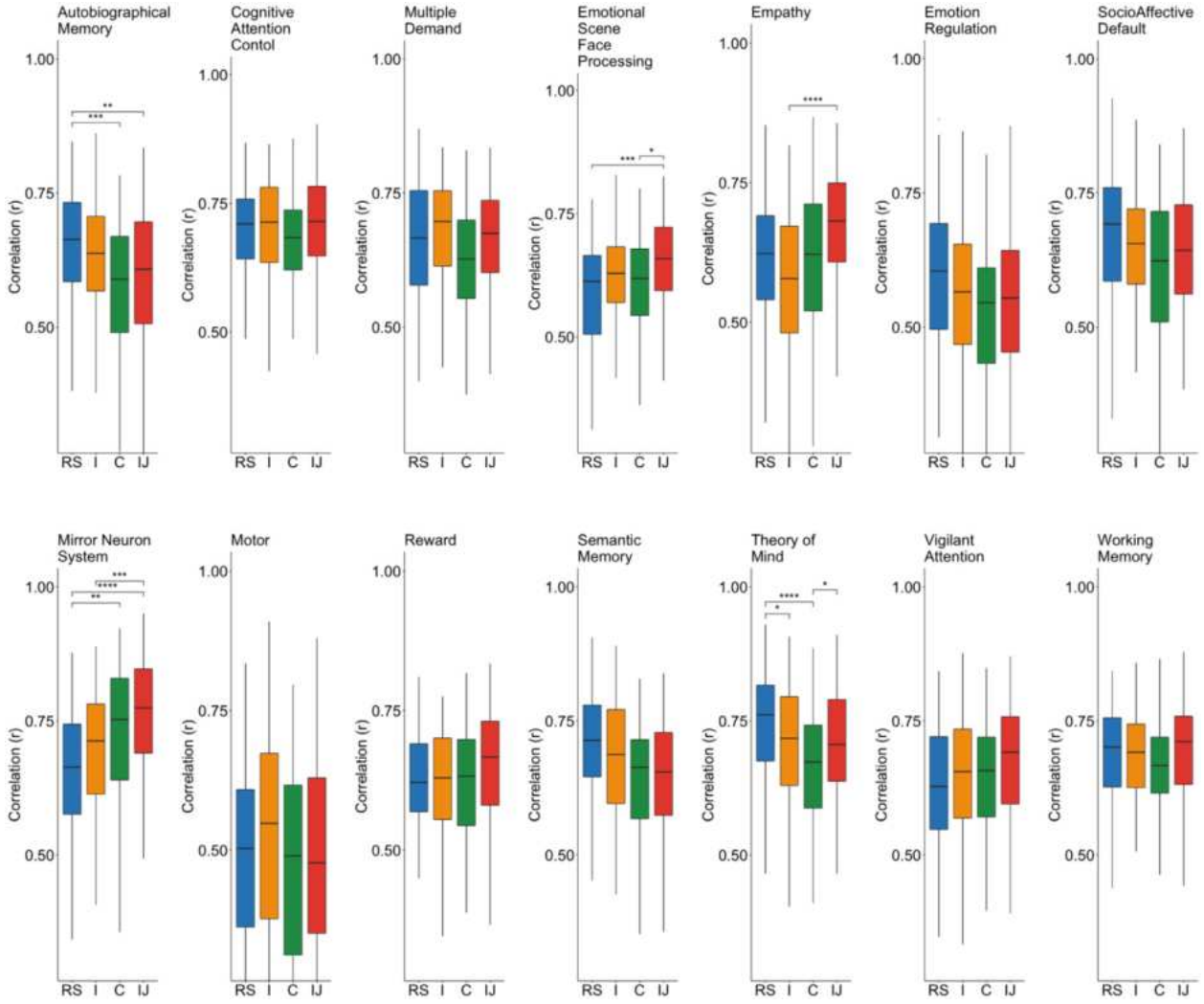


Fig. 4. Within-subject correlations for the meta-analytically defined networks. Correlations across all session pairings are depicted. (RS= Resting State, *I* = Inscapes, C = Circus, IJ = Indiana Jones).

more, as the motor network was created solely based on fingertapping tasks, it seems reasonable to assume that activation was low in this network. Therefore, connectivity patterns are expected to be rather similar across participants.

Indiana Jones was the stimulus that achieved the highest identification accuracy in the majority of networks. Previous studies have argued that the major driving factor for improvement of individual identifiability is the social content of a stimulus (Nummenmaa et al., 2014; Schmälzle et al., 2015; Dmochowski et al., 2014)., which in the present study was most pronounced for *Indiana Jones*. In comparison, neither *Circus* nor *Inscapes* reach the level of social content depicted in *Indiana Jones. Circus'* complete lack of speech might have taken away from the social component whereas *Inscapes* does not depict any human interaction at all.

4.3. Identification accuracies for different network sizes

Since we observed an increase of identification accuracy with network size such that bigger networks tended to show higher accuracies, we investigated the influence of network size on identifiability in artificially created networks (Fig. 3). The results show the same tendency that was observed in the meta-analytically defined networks, such that identification accuracy was highest for *Indiana Jones*, followed by RS, *Inscapes* and *Circus*. Confirming our observation, identification accuracy in artificial networks increased with network size, regardless of condition. Notably, all meta-analytical networks, except the motor network, outperformed artificial networks of the same size, supporting their biological validity. Following our previous line of argument, the motor network might not be suitable for subject identification based on FC, which might explain the underperformance compared to artificial networks.

4.4. Within- and between-subject correlations in meta-analytic networks

To better understand the differences in identifiability across stimuli and networks, we investigated within- and between-subject correlations (Figs. 4 and 5). Our results showed that in the majority of networks, within- and between-subject correlations were significantly altered during NV in comparison to RS. It is generally assumed that NV should increase between-subject similarity, given that all subjects are presented with the same stimuli, in comparison to no stimuli at all during RS (Hasson et al., 2004; Hasson et al., 2010; Kauppi, 2010). On the other hand, it is unclear whether NV can evoke unique and reliable patterns across sessions, as measured by within-subject correlations. Vanderwal and colleagues investigated FC variability in NV and RS and showed that naturalistic paradigms increased within- and between-subject correlations on a whole brain level (Vanderwal et al., 2017). However, our results showed no significant differences for either within- or between-

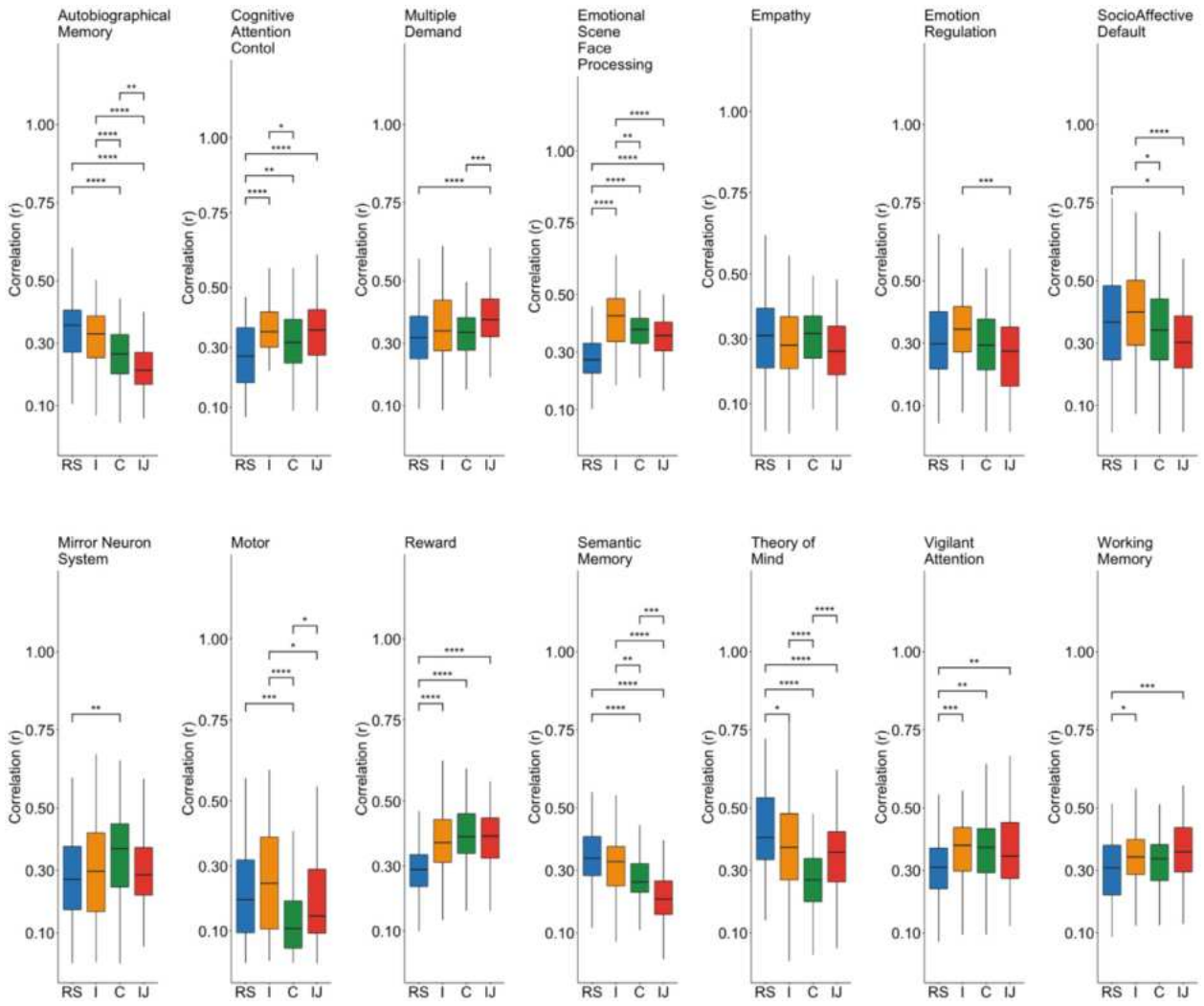


Fig. 5. Between-subject correlations for the meta-analytically defined networks. Correlations for all sessions are depicted. (RS= Resting State, *I* = Inscapes, C = Circus, IJ = Indiana Jones).

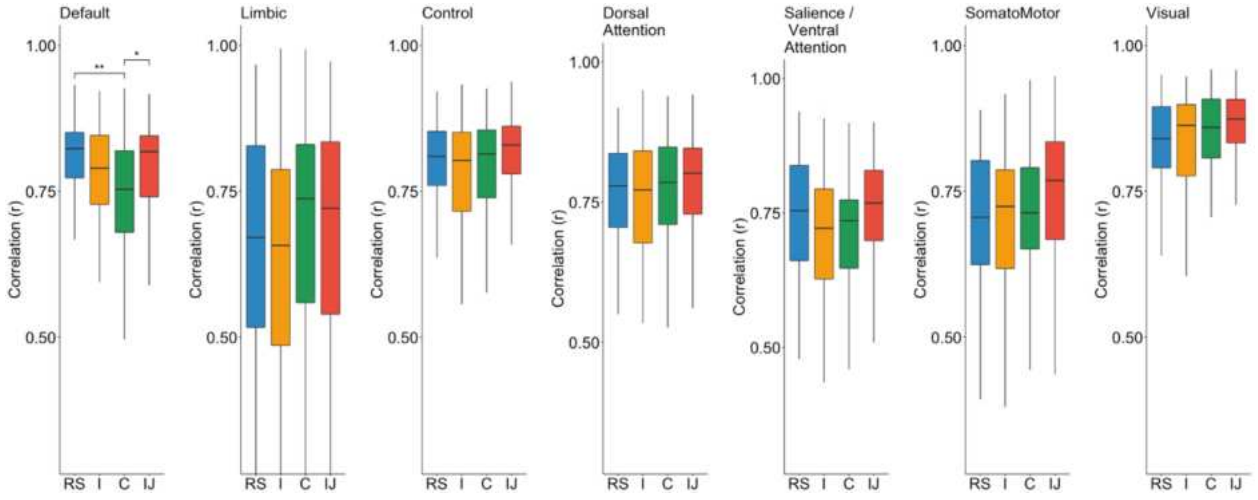


Fig. 6. Within-subject correlations for the RS derived networks. Correlations across all session pairings are depicted.

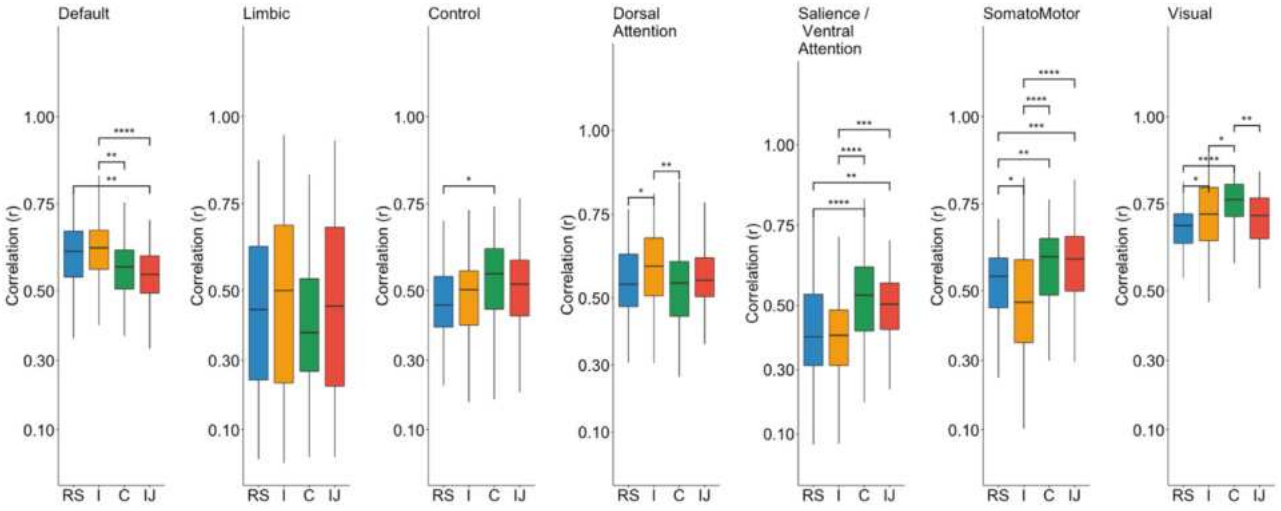


Fig. 7. Between-subject correlations for the RS derived networks. Correlations for all sessions are depicted.

subject correlations on a whole-brain level (supplementary Fig. S1 and S2). On the other hand, our analysis revealed varying effects on a network basis. Increased within-subject correlations were mainly observed in meta-analytic networks that are essential for perception and processing of action, behavior and emotions, namely EmoSF, Empathy and MNS. In other networks, NV resulted in more similar patterns between subjects (CogAC, eMDN, Rew, VigAtt and WM). Notably, multiple networks showed decreased within- and between-subject correlations during NV (AM, ER, eSAD, SM, ToM). We will discuss these three groups of networks subsequently.

4.4.1. Networks with higher within-subject correlations in movies

NV showed significantly higher within-subject correlations in networks that are essential for perception and processing of action, behavior and emotions (EmoSF and MNS). In a recent publication by Finn and Bandettini (2020) it was shown that NV outperformed RS in FC-based prediction of behavioral scores. In their study, movies with strong social content led to the more accurate predictions, regardless of whether the predicted score was social or cognitive. The authors hypothesize that social movies are not only more engaging, but also most likely to evoke divergent interpretations and reactions across participants. In agreement with this assumption, several studies have shown that social movies induce different neural responses across subjects (Finn et al., 2018; Rikandi et al., 2017; Salmi et al., 2020) and that shared interpretation of a narrative or movie is associated with similarity in neural responses (Nguyen et al., 2019; Gruskin et al., 2020). Assuming that the social aspect of a movie stimulus induces stable individual connectivity patterns, it is reasonable to expect that this effect is more pronounced in networks that deal with the processing of social interactions.

In the EmoSF network for example, which deals with the visual and emotional processing of faces or scenes (Sabatinelli et al., 2011), all three movie stimuli led to higher within-subject correlations compared to RS. Notably, the movie *Indiana Jones*, during which the emotional processing of faces is a key aspect, shows highest within-subject correlations. Here, differences in the emotional assessment of the particular faces and scenes might have been the driving factor that evoked stable individual connectivity patterns during *Indiana Jones*.

In the MNS network, which is involved in the understanding of actions and their underlying intentions as well as the imitation of observed behavior, we observed an increase of individuality with increasingly complex stimuli (Caspers et al., 2010). Especially the two stimuli *Circus* and *Indiana Jones*, during which action and behavior of different characters are depicted, should engage the MNS network which in turn might have led to the increased within-subject correlations. The between-subject correlations were significantly stronger for *Circus* than for *RS*, but not different between the remaining conditions. Presumably, the movie *Circus* serves as the optimal stimulus for action observation since it shows moving characters, but (unlike *Indiana Jones*) does not include competing stimuli like speech.

Another network that showed similar patterns, although not reaching significance is the Empathy network, which deals with the emotional cognition of moral behavior (Bzdok et al., 2012). The within-subject correlations were increased during the movies *Circus* and *Indiana Jones*. During both movies, characters show varying emotions in response to different situations, which might have been experienced differently across subjects. *Inscapes* on the other hand performed similar to RS, likely because the depicted abstract shapes failed to engage the network.

4.4.2. Networks with higher between-subject correlations in movies

NV showed significantly higher between-subject correlations in networks that are associated with executive functions and/or stimulus evaluation (CogAC, eMDN, EmoSF, MNS, Rew, VigAtt and WM). Here, NV increased the similarity of FC across participants (i.e. higher-between subject correlation), but did not increase within-subject correlations. Several other studies have found NV to increase the similarity between subjects (Finn et al., 2020; Hasson et al., 2004; Vanderwal et al., 2017; Wang et al., 2017), which is likely caused by exposure to the same stimulus. Although these studies mostly agree that individual differences can exist on top of the shared response on a whole-brain level, they acknowledge two possible scenarios: On the one hand, the stimulus evoked similarity across subjects might enable better observation of individual differences (Vanderwal et al., 2017; Finn and Bandettini, 2020). On the other hand, strongly increased similarity across subjects' neuronal response might blur individual features (Finn et al., 2017). The same assumptions hold true from a network perspective, such that networks subjected to the same stimulus can either exhibit deviating patterns on top of the shared response, or highly similar patterns which conceal individual differences, depending on the specific network function.

Considering the main function of each respective network, none of the networks should be particularly engaged during RS or during any of the movies. The CogAC network is essential for the suppression of a predominant but inadequate response in favor of the contextually appropriate response (Cieslik et al., 2015). The eMDN consists of core regions that are active during most processes which involve executive or higher cognitive functions and a set of more task-specific regions extending these core regions (Camilleri et al., 2018). The Rew network is essential for reward-related decision making (Liu et al., 2011). The

Table 2
Identification accuracies per network and modality, averaged across sessions. Networks are in order of highest average accuracy. The highest identification accuracy in each network is denoted in bold.

Network	RS	Inscapes	Circus	Indiana Jones	Node Number
Default	**91.18%**	85.78%	87.75%	90.69%	24
Control	86.76%	87.75%	79.90%	**89.71%**	13
Dorsal Attention	75.49%	73.04%	66.18%	**75.98%**	15
Salience	**74.51%**	66.67%	56.37%	73.04%	12
Visual	68.63%	58.82%	57.84%	**73.04%**	17
Somatomotor	**62.75%**	60.29%	43.14%	57.84%	14
Limbic	14.22%	14.22%	9.31%	12.25%	5

VigAtt network is involved in vigilant attention, i.e. the continued focus on intellectually un-challenging tasks (Langner and Eickhoff, 2013). The WM network is fundamental for the storage and manipulation of short-term memory (Rottschy et al., 2012). Since individual differences are likely only enhanced in networks that are engaged during a certain condition, we assume that NV did not evoke stable individual connectivity patterns, as the processing of movies may not rely on the core network function. Therefore, subjects are less unique and more similar to themselves, increasing between-subject correlations especially in comparison with unconstrained RS where more heterogeneous responses are expected.

4.4.3. Networks with higher between- or within-subject correlations in RS

The vast majority of previous studies reported increased within- and between-subject correlations for NV in comparison with RS (Finn et al., 2020; Hasson et al., 2004; Vanderwal et al., 2017; Wang et al., 2017; Nastase et al., 2019). However, all of these studies employed analyses of whole brain connectivity, disregarding effects in single networks. While previous result patterns hold true in some networks, we also show that NV decreases within- and between-subject correlations in other networks (AM, ER, eSAD, SM and ToM).

The majority of these networks at least partially overlap with the default mode network, which is tied to intrinsically oriented functions, rather than the processing of external stimuli (Hasson et al., 2004; Golland et al., 2007). Therefore, it seems plausible that NV does not increase within- or between-subject correlations in these networks which are likely not engaged during movie watching. The AM network was the only network in which within-subject correlations for RS exceeded *Indiana Jones*. This network comprises brain regions engaged in processes concerning scene-construction and self-projection, or the ability to mentally project oneself from the present moment into another time, place, or perspective. Consequently we would expect the AM network to be more strongly activated during RS, when the mind is not occupied by the content of a movie. Our data indeed shows that participants during RS showed higher within-subject correlations than during the two narrative movie clips *Circus* and *Indiana Jones*, but not significantly different from the purely abstract animation *Inscapes*. Therefore, we conclude that in absence of a storyline, subjects divert to imagined situations instead of the external stimuli, thus engaging the AM network which leads to higher within-subject correlations for RS than for the narrative movies. We assume that the movie *Inscapes* is inbetween a narrative and the complete absence of a stimuli, thus it may fail to engage participants over a longer period of time, therefore letting the participant zone out eventually. In addition, RS and *Inscapes* also increased between-subject correlations in comparison to both narrative movies. Likely, increased between-subject correlations are driven by the joint activation of the AM network during RS and *Inscapes*. On the other hand, *Circus* and *Indiana Jones* likely engage the network to a lesser extent, thereby falling short of evoking coordinated activity which in turn reduces similarity between subjects.

The eSAD network was defined to comprise those brain regions that are part of the default mode network, but at the same time also involved in social or affective processing (Amft et al., 2015). Thus, the network is engaged in socio-affective processing including emotional processes, cognition, reward, introception, memory and theory of mind functions. Although not exclusively a “task-negative” network, the eSAD network is highly overlapping with the default mode network and generally presumed to be more active when participants can let their thoughts run free (Amft et al., 2015). RS showed higher within-subject correlations than *Circus* as well as higher between-subject correlations than *Indiana Jones*. In addition, *Inscapes*, which is arguably closer to RS than the other movies, also showed higher between-subject correlations as compared to *Circus* and *Indiana Jones*. Due to the default mode aspects of the eSAD network, it is perceivable that this network is more strongly engaged during RS and *Inscapes*. Thus, participants are more likely to express different connectivity patterns as compared to NV where the network is mostly unengaged. The movies *Circus* and *Indiana Jones* on the other hand might result in a less pronounced engagement of the network, thus failing to evoke similar patterns across participants.

The SM network is involved in retrieving semantic knowledge and is highly overlapping with the default mode network (Binder et al., 2009). The authors argue that task-unrelated thoughts are inherently semantic, because they require the manipulation of stored knowledge (Binder et al., 1999). Furthermore, semantic processing was reliably shown to be suppressed during demanding perceptual tasks (Binder et al., 2009), which is in accordance with our result pattern, showing increasingly complex stimuli to decrease within- and between-subject similarity (RS > *Inscapes* > *Circus* > *Indiana Jones*). We thus suggest that increasing complexity of the movie stimuli suppresses semantic processing and therefore leads to less engagement of the SM network. Presumably, due to a less pronounced engagement of the SM network during that *Circus* and *Indiana Jones*, participants show low between-subject correlations as well as low within-subject correlations.

The ToM network is fundamental for the understanding and contemplation of the behavior and intentions of others (Bzdok et al., 2012). Within- and between-subject correlations in the ToM network were generally higher during RS than during the NV conditions. We assume that movies evoke different interpretations of the intentions of the depicted characters and thus may have led to diverging connectivity profiles, in turn increasing differences between subjects. On the other hand, these differences seem to be unstable across sessions, thus decreasing within-subject correlations during NV.

4.5. Comparison with RS-derived networks

Identification accuracies in RS-derived networks confirm the assumption that identifiability is dependent on the network-stimulus combination (Table 2). Highest identification accuracy for RS was achieved in the Default, Salience and SomatoMotor networks, whereas highest accuracy for *Indiana Jones* was found in the Control, Dorsal Attention and Visual networks. For RS highest overall accuracy (91%) was achieved in the Default network, which is prominently active during RS (Long et al., 2008). However, the accuracies achieved in RS-derived networks were generally lower than those achieved in meta-analytic networks. Out of

the 14 meta-analytic networks, eight yielded higher accuracies than the best performing RS derived network (Default).

In accordance with our results on meta-analytic networks, within- and between-subject correlations were also significantly altered during NV, in comparison to RS, in the RS-derived networks (Figs. 6 and 7). In the Control, Dorsal Attention, Salience, SomatoMotor and Visual networks NV resulted in more similar patterns between subjects. Only in the Default network, NV showed decreased between-subject correlations in comparison with RS.

Noticeably, differences in within-subject correlations between NV conditions and RS are less pronounced in the RS-derived networks than in the meta-analytic networks. This is further supported by the fact that RDMs of RS and NV stimuli tended to cluster together more often in RS derived networks (Fig. 2). Furthermore, within- and especially between-subject correlations are largely increased for the RS networks, resulting in reduced identifiability in RS derived networks compared to the meta-analytic networks. On the one hand, meta-analytic networks seem to be more sensitive to differences between NV stimuli and RS, likely because they best represent the core nodes of a given cognitive function. On the other hand, although within-subject correlations are increased in RS derived networks, the larger increase in between-subject similarity overshadows this effect and consequently leads to decreased identifiability. Taken together, the present results underline the viability of using specific meta-analytic networks for reliably identifying subjects' brain connectivity patterns under NV conditions.

4.6. Limitations

While the current study sheds new light onto individual differences in, and stability of, brain states elicited by movie watching, it comes with some limitations. Firstly, individual outliers might have biased identification accuracies, due to the small sample size. However, previous studies on RS and NV reported similar identification accuracies as those achieved in this study (Finn et al., 2015; Vanderwal et al., 2017). Nevertheless, future studies should be conducted on larger samples to confirm our results. Secondly, while we demonstrated enhanced individual difference and identifiability for certain stimulus-network combinations, our study did not include any phenotypes. Therefore this study is not suited to determine whether enhanced individual differences under NV can be used to more accurately predict phenotypes as compared to RS. Hence, future studies should investigate the interplay between increased identifiability and the accuracy of phenotype predictions. Thirdly, reliability of FC might at least partly be driven by structured noise such as vascular effects (Varikuti et al., 2017). Although we applied a number of denoising strategies, results might thus be confounded by non-neuronal signals. Additionally, only static FC was considered in the present study. Future studies investigating dynamic FC might shed more light on how individual variability in functional brain organization changes over the time course of a movie. A previous study on dynamic FC showed that NV improved test-retest reliability over RS, similar to the results in this study (Zhang et al., 2022). Finally, it was not assessed whether participants had seen any of the movie clips prior to participating in the study. Knowing the film beforehand could affect engagement of the participant and thereby modulate the effect of NV. In addition, previous studies have shown that expected stimuli can decrease the neuronal response (Alink et al., 2010; Koster-Hale and Saxe, 2013). Since the three sessions in our study were conducted within a week, participants are expected to be rather familiar with the movie content during the second and third session. Therefore, it is possible that the predictable content reduced the neuronal response and influenced our results. However, our results showed that connectivity patterns rather clustered according to stimulus than repetition, which suggests that the same movie stimulus can be used repeatedly to study FC of a subject across various time points. Similarly, a study by Wang et al. (2017) showed that movie fMRI increased reliability over RS across two sessions. The authors concluded that the effect that is achieved by increased engagement during movie watching, outweighs the impact of familiarity with a given movie. Taken together, these findings encourage the application of movie fMRI in clinical studies where it is necessary to monitor patients over a longer period of time.

5. Conclusions

NV has been suggested to show high potential for emphasizing individual differences, but effects have often been reported on a whole-brain level only. Our study extends the current knowledge by characterizing the influence of NV on FC in meta-analytically derived functional networks. We show that NV increases identifiability of individuals based on functional connectivity in certain networks. However, there is not one naturalistic stimulus that will enhance individual differences across the brain. Therefore it is crucial to select the appropriate stimulus and networks for the research question at hand.

Declaration of Competing Interest

The authors report no competing interests.

Credit authorship contribution statement

Jean-Philippe Kröll: Formal analysis, Software, Validation, Visualization, Writing – original draft. **Patrick Friedrich:** Writing – review & editing, Methodology. **Xuan Li:** Writing – review & editing. **Kaustubh R. Patil:** Writing – review & editing. **Lisa Mochalski:** Writing – review & editing. **Laura Waite:** Data curation, Writing – review & editing. **Xing Qian:** Writing – review & editing. **Michael WL Chee:** Writing – review & editing. **Juan Helen Zhou:** Conceptualization, Writing – review & editing, Funding acquisition. **Simon Eickhoff:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **Susanne Weis:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Data and code availability statement

The data that support the findings of this study are available upon request with data sharing agreement from the co-author Dr. Helen Zhou, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon request and with permission of the Institutional Review Board of the National University of Singapore.

Custom code used in this study is available upon request from the corresponding author JK.

Acknowledgements

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 945539 (HBP SGA3), and the Deutsche Forschungsgemeinschaft (491111487). We also acknowledge the funding support from Yong Loo Lin School of Medicine, National University of Singapore (J.H.Z), the Duke-NUS Medical School Signature Research Program Core Funding (J.H.Z.), and Ministry of Education, Singapore (MOE-T2EP40120–0007, J.H.Z), and Far East Organization (E-546–00–0398–01, MWLC.)

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120083.

References

Alink, A., Schwiedrzik, C.M., Kohler, A., Singer, W., Muckli, L., 2010. Stimulus predictability reduces responses in primary visual cortex. J. Neurosci. 30 (8), 2960–2966. doi:10.1523/JNEUROSCI.3730-10.2010.

Amft, M., Bzdok, D., Laird, A.R., Fox, P.T., Schilbach, L., Eickhoff, S.B., 2015. Definition and characterization of an extended social-affective default network. Brain Struct. Funct. 220 (2), 1031–1049. doi:10.1007/s00429-013-0698-0.

Avants, B., Tustison, N.J., Song, G., 2009. Advanced normalization tools: V1.0. Insight J. doi:10.54294/uvnhin, Published online July 29.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Rao, S.M., Cox, R.W., 1999. Conceptual processing during the conscious resting state: a functional MRI study. J. Cogn. Neurosci. 11 (1), 80–93. doi:10.1162/089892999563265.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex 19 (12), 2767–2796. doi:10.1093/cercor/bhp055.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Reson. Med. 34 (4), 537–541. doi:10.1002/mrm.1910340409.

Buhle, J.T., Silvers, J.A., Wager, T.D., et al., 2014. Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. Cereb. Cortex 24 (11), 2981–2990. doi:10.1093/cercor/bht154.

Bzdok, D., Schilbach, L., Vogeley, K., et al., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct. Funct. 217 (4), 783–796. doi:10.1007/s00429-012-0380-y.

Camilleri, J.A., Müller, V.I., Fox, P., et al., 2018. Definition and characterization of an extended multiple-demand network. Neuroimage 165, 138–147. doi:10.1016/j.neuroimage.2017.10.020.

Caspers, S., Zilles, K., Laird, A.R., Eickhoff, S.B., 2010. ALE meta-analysis of action observation and imitation in the human brain. Neuroimage 50 (3), 1148–1167. doi:10.1016/j.neuroimage.2009.12.112.

Christoff, K., Ream, J.M., Gabrieli, J.D.E., 2004. Neural basis of spontaneous thought processes. Cortex 40 (4–5), 623–630. doi:10.1016/S0010-9452(08)70158-8.

Cieslik, E.C., Mueller, V.I., Eickhoff, C.R., Langner, R., Eickhoff, S.B., 2015. Three key regions for supervisory attentional control: evidence from neuroimaging meta-analyses. Neurosci. Biobehav. Rev. 48, 22–34. doi:10.1016/j.neubiorev.2014.11.003.

Croxson, P.L., Forkel, S.J., Cerliani, L., Thiebaut de Schotten, M, 2018. Structural variability across the primate brain: a cross-species comparison. Cereb. Cortex 28 (11), 3829–3841. doi:10.1093/cercor/bhx244.

Damoiseaux, J.S., Rombouts, S.A.R.B., Barkhof, F., et al., 2006. Consistent resting-state networks across healthy subjects. Proc. Natl. Acad. Sci. 103 (37), 13848–13853. doi:10.1073/pnas.0601417103.

van der Meer, J.N., Breakspear, M., Chang, L.J., Sonkusare, S., Cocchi, L, 2020. Movie viewing elicits rich and reliable brain state dynamics. Nat. Commun. 11 (1), 5004. doi:10.1038/s41467-020-18717-w.

Dmochowski, J.P., Bezdek, M.A., Abelson, B.P., Johnson, J.S., Schumacher, E.H., Parra, L.C., 2014. Audience preferences are predicted by temporal reliability of neural processing. Nat. Commun. 5 (1), 4567. doi:10.1038/ncomms5567.

Dosenbach, N.U.F., Fair, D.A., Miezin, F.M., et al., 2007. Distinct brain networks for adaptive and stable task control in humans. Proc. Natl. Acad. Sci. U. S. A. 104 (26), 11073–11078. doi:10.1073/pnas.0704320104.

Eickhoff, S.B., Bzdok, D., Laird, A.R., Kurth, F., Fox, P.T., 2012. Activation likelihood estimation meta-analysis revisited. Neuroimage 59 (3), 2349–2361. doi:10.1016/j.neuroimage.2011.09.017.

Eickhoff, S.B., Milham, M., Vanderwal, T., 2020. Towards clinical applications of movie fMRI. Neuroimage 217, 116860. doi:10.1016/j.neuroimage.2020.116860.

Esteban, O., Markiewicz, C.J., Blair, R.W., et al., 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods 16 (1), 111–116. doi:10.1038/s41592-018-0235-4.

Finn, E.S., Bandettini, P.A., 2020. Movie-watching outperforms rest for functional connectivity-based prediction of behavior. Neuroscience doi:10.1101/2020.08.23.263723.

Finn, E.S., Shen, X., Scheinost, D., et al., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18 (11), 1664–1671. doi:10.1038/nn.4135.

Finn, E.S., Scheinost, D., Finn, D.M., Shen, X., Papademetris, X., Constable, R.T., 2017. Can brain state be manipulated to emphasize individual differences in functional connectivity? Neuroimage 160, 140–151. Accessed June 29, 2022 https://linkinghub.elsevier.com/retrieve/pii/S1053811917302872.

Finn, E.S., Corlett, P.R., Chen, G., Bandettini, P.A., Constable, R.T., 2018. Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. Nat. Commun. 9 (1), 2043. doi:10.1038/s41467-018-04387-2.

Finn, E.S., Glerean, E., Khojandi, A.Y., et al., 2020. Idiosynchrony: from shared responses to individual differences during naturalistic neuroimaging. Neuroimage 215, 116828. doi:10.1016/j.neuroimage.2020.116828.

Fox, M.D., Corbetta, M., Snyder, A.Z., Vincent, J.L., Raichle, M.E., 2006. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. Proc. Natl. Acad. Sci. U. S. A. 103 (26), 10046–10051. doi:10.1073/pnas.0604187103.

Geerligs, L., Rubinov, M., Cam-CAN, Henson, RN, 2015. State and trait components of functional connectivity: individual differences vary with mental state. J. Neurosci. 35 (41), 13949–13961. doi:10.1523/JNEUROSCI.1324-15.2015.

Golland, Y., Bentin, S., Gelbard, H., et al., 2007. Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. Cereb. Cortex 17 (4), 766–777. doi:10.1093/cercor/bhk030.

Gonzalez-Castillo, J., Kam, J.W.Y., Hoy, C.W., Bandettini, P.A., 2021. How to interpret resting-state fMRI: ask your participants. J. Neurosci. 41 (6), 1130–1141. doi:10.1523/JNEUROSCI.1786-20.2020.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proc. Natl. Acad. Sci. U. S. A. 100 (1), 253–258. doi:10.1073/pnas.0135058100.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48 (1), 63–72. doi:10.1016/j.neuroimage.2009.06.060.

Gruskin, D.C., Rosenberg, M.D., Holmes, A.J., 2020. Relationships between depressive symptoms and brain responses during emotional movie viewing emerge in adolescence. Neuroimage 216, 116217. doi:10.1016/j.neuroimage.2019.116217.

Halchenko, Y., Meyer, K., Poldrack, B., et al., 2021. DataLad: distributed system for joint management of code, data, and their relationship. JOSS 6 (63), 3262. doi:10.21105/joss.03262.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303 (5664), 1634–1640. doi:10.1126/science.1089506.

Hasson, U., Malach, R., Heeger, D.J., 2010. Reliability of cortical activity during natural stimulation. Trends Cogn. Sci. (Regul. Ed.) 14 (1), 40–48. doi:10.1016/j.tics.2009.10.011.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17 (2), 825–841. doi:10.1006/nimg.2002.1132.

Kauppi, 2010. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Front. Neuroinform doi:10.3389/fninf.2010.00005, Published online.

Koster-Hale, J., Saxe, R., 2013. Theory of mind: a neural prediction problem. Neuron 79 (5), 836–848. doi:10.1016/j.neuron.2013.08.020.

Kriegeskorte, N., 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. doi:10.3389/neuro.06.004.2008, Published online.

Langner, R., Eickhoff, S.B., 2013. Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. Psychol. Bull. 139 (4), 870–900. doi:10.1037/a0030694.

Liu, X., Hairston, J., Schrier, M., Fan, J., 2011. Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. Neurosci. Biobehav. Rev. 35 (5), 1219–1236. doi:10.1016/j.neubiorev.2010.12.012.

Long, X.Y., Zuo, X.N., Kiviniemi, V., et al., 2008. Default mode network as revealed with multiple methods for resting-state functional MRI analysis. J. Neurosci. Methods 171 (2), 349–355. doi:10.1016/j.jneumeth.2008.03.021.

McInnes, L., Healy, J., Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Published online 2018. doi:10.48550/ARXIV.1802.03426.

Mennes, M., Kelly, C., Zuo, X.N., et al., 2010. Inter-individual differences in resting-state functional connectivity predict task-induced BOLD activity. Neuroimage 50 (4), 1690–1701. doi:10.1016/j.neuroimage.2010.01.002.

Mishra, S., Srinivasan, N., Tiwary, U.S., 2022. Dynamic functional connectivity of emotion processing in beta band with naturalistic emotion stimuli. Brain Sci. 12 (8), 1106. doi:10.3390/brainsci12081106.

Nastase, S.A., Gazzola, V., Hasson, U., Keysers, C., 2019. Measuring shared responses across subjects using intersubject correlation. Soc. Cogn. Affect. Neurosci. nsz037. doi:10.1093/scan/nsz037, Published online May 16.

Nguyen, M., Vanderwal, T., Hasson, U., 2019. Shared understanding of narratives is correlated with shared neural responses. Neuroimage 184, 161–170. doi:10.1016/j.neuroimage.2018.09.010.

Nummenmaa, L., Saarimäki, H., Glerean, E., et al., 2014. Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks. Neuroimage 102, 498–509. doi:10.1016/j.neuroimage.2014.07.063.

Patriat, R., Molloy, E.K., Meier, T.B., et al., 2013. The effect of resting condition on resting-state fMRI reliability and consistency: a comparison between resting with eyes open, closed, and fixated. Neuroimage 78, 463–473. doi:10.1016/j.neuroimage.2013.04.013.

Power, J.D., Cohen, A.L., Nelson, S.M., et al., 2011. Functional network organization of the human brain. Neuron 72 (4), 665–678. doi:10.1016/j.neuron.2011.09.006.

Pruim, R.H.R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J.K., Beckmann, C.F., 2015. ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. Neuroimage 112, 267–277. doi:10.1016/j.neuroimage.2015.02.064.

Rikandi, E., Pamilo, S., Mäntylä, T., et al., 2017. Precuneus functioning differentiates first-episode psychosis patients during the fantasy movie Alice in Wonderland. Psychol. Med. 47 (3), 495–506. doi:10.1017/S0033291716002609.

Rottschy, C., Langner, R., Dogan, I., et al., 2012. Modelling neural correlates of working memory: a coordinate-based meta-analysis. Neuroimage 60 (1), 830–846. doi:10.1016/j.neuroimage.2011.11.050.

Saarimäki, H., 2021. Naturalistic stimuli in affective neuroimaging: a review. Front. Hum. Neurosci. 15, 675068. doi:10.3389/fnhum.2021.675068.

Sabatinelli, D., Fortune, E.E., Li, Q., et al., 2011. Emotional perception: meta-analyses of face and natural scene processing. Neuroimage 54 (3), 2524–2533. doi:10.1016/j.neuroimage.2010.10.011.

Salmi, J., Metwaly, M., Tohka, J., et al., 2020. ADHD desynchronizes brain activity during watching a distracted multi-talker conversation. Neuroimage 216, 116352. doi:10.1016/j.neuroimage.2019.116352.

Schaefer, A., Nils, F., Sanchez, X., Philippot, P., 2010. Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. Cogn. Emot. 24 (7), 1153–1172. doi:10.1080/02699930903274322.

Schaefer, A., Kong, R., Gordon, E.M., et al., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb. Cortex 28 (9), 3095–3114. doi:10.1093/cercor/bhx179.

Schmälzle, R., Häcker, F.E.K., Honey, C.J., Hasson, U., 2015. Engaged listeners: shared neural processing of powerful political speeches. Soc. Cogn. Affect. Neurosci. 10 (8), 1137–1143. doi:10.1093/scan/nsu168.

Shehzad, Z., Kelly, A.M.C., Reiss, P.T., et al., 2009. The Resting Brain: unconstrained yet Reliable. Cereb. Cortex 19 (10), 2209–2229. doi:10.1093/cercor/bhn256.

Shen, X., Tokoglu, F., Papademetris, X., Constable, R.T., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. Neuroimage 82, 403–415. doi:10.1016/j.neuroimage.2013.05.081.

Smith, S.M., Fox, P.T., Miller, K.L., et al., 2009. Correspondence of the brain's functional architecture during activation and rest. Proc. Natl. Acad. Sci. U. S. A. 106 (31), 13040–13045. doi:10.1073/pnas.0905267106.

Sporns, O., Betzel, R.F., 2016. Modular Brain Networks. Annu. Rev. Psychol. 67 (1), 613–640. doi:10.1146/annurev-psych-122414-033634.

Spreng, R.N., Mar, R.A., Kim, A.S.N., 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J. Cogn. Neurosci. 21 (3), 489–510. doi:10.1162/jocn.2008.21029.

Thomas Yeo, B.T., Krienen, F.M., Sepulcre, J., et al., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106 (3), 1125–1165. doi:10.1152/jn.00338.2011.

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L.C., Castellanos, F.X., 2015. Inscapes : a movie paradigm to improve compliance in functional magnetic resonance imaging. Neuroimage 122, 222–232. doi:10.1016/j.neuroimage.2015.07.069.

Vanderwal, T., Eilbott, J., Finn, E.S., Craddock, R.C., Turnbull, A., Castellanos, F.X., 2017. Individual differences in functional connectivity during naturalistic viewing conditions. Neuroimage 157, 521–530. doi:10.1016/j.neuroimage.2017.06.027.

Vanderwal, T., Eilbott, J., Castellanos, F.X., 2019. Movies in the magnet: naturalistic paradigms in developmental functional neuroimaging. Dev. Cogn. Neurosci. 36, 100600. doi:10.1016/j.dcn.2018.10.004.

Varikuti, D.P., Hoffstaedter, F., Genon, S., Schwender, H., Reid, A.T., Eickhoff, S.B., 2017. Resting-state test–retest reliability of a priori defined canonical networks over different preprocessing steps. Brain Struct. Funct. 222 (3), 1447–1468. doi:10.1007/s00429-016-1286-x.

Wang, J., Ren, Y., Hu, X., et al., 2017. Test-retest reliability of functional connectivity networks during naturalistic fMRI paradigms: test-retest reliability of naturalistic fMRI. Hum. Brain Mapp. 38 (4), 2226–2241. doi:10.1002/hbm.23517.

Witt, S.T., Laird, A.R., Meyerand, M.E., 2008. Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. Neuroimage 42 (1), 343–356. doi:10.1016/j.neuroimage.2008.04.025.

Zhang, X., Liu, J., Yang, Y., et al., 2022. Test–retest reliability of dynamic functional connectivity in naturalistic paradigm functional magnetic resonance imaging. Hum. Brain Mapp. 43 (4), 1463–1476. doi:10.1002/hbm.25736.

4 Inter- and intra-subject similarity in network functional connectivity across a full narrative movie, Lisa N. Mochalski, Patrick Friedrich, Xuan Li, Jean-Philippe Kröll, Simon B. Eickhoff, Susanne Weis, *Human Brain Mapping*, *45*(11), e26802, (2024)

Received: 27 June 2023 | Revised: 11 July 2024 | Accepted: 17 July 2024

DOI: 10.1002/hbm.26802

RESEARCH ARTICLE

WILEY

Inter- and intra-subject similarity in network functional connectivity across a full narrative movie

Lisa N. Mochalski[1,2] | Patrick Friedrich[1] | Xuan Li[1] | Jean-Philippe Kröll[1] | Simon B. Eickhoff[1,2] | Susanne Weis[1,2]

[1]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany

[2]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Correspondence
Susanne Weis, Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich 52428, Germany.
Email: s.weis@fz-juelich.de

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: 945539; Deutsche Forschungsgemeinschaft (DFG), Grant/Award Number: 491111487; National Institute of Mental Health, Grant/Award Number: R01-MH074457

Abstract

Naturalistic paradigms, such as watching movies during functional magnetic resonance imaging, are thought to prompt the emotional and cognitive processes typically elicited in real life situations. Therefore, naturalistic viewing (NV) holds great potential for studying individual differences. Previous studies have primarily focused on using shorter movie clips, geared toward eliciting specific and often isolated emotions, while the potential behind using full narratives depicted in commercial movies as a proxy for real-life experiences has barely been explored. Here, we offer preliminary evidence that a full narrative movie (FNM), that is, a movie covering a complete narrative arc, can capture complex socio-affective dynamics and their links to individual differences. Using the studyforrest dataset, we investigated inter- and intra-subject similarity in network functional connectivity (NFC) of 14 meta-analytically defined networks across a full narrative, audio-visual movie split into eight consecutive movie segments. We characterized the movie segments by valence and arousal portrayed within the sequences, before utilizing a linear mixed model to analyze which factors explain inter- and intra-subject similarity. Our results show that the model best explaining inter-subject similarity comprised network, movie segment, valence and a movie segment by valence interaction. Intra-subject similarity was influenced significantly by the same factors and an additional three-way interaction between movie segment, valence and arousal. Overall, inter- and intra-subject similarity in NFC were sensitive to the ongoing narrative and emotions in the movie. We conclude that FNMs offer complex content and dynamics that might be particularly valuable for studying individual differences. Further characterization of movie features, such as the overarching narratives, that enhance individual differences is needed for advancing the potential of NV research.

KEYWORDS
individual differences, meta-analytical networks, movie fMRI, naturalistic viewing, network functional connectivity

1 | INTRODUCTION

Understanding how individual differences in brain architecture shape personality, cognitive abilities and socio-affective traits is a constant endeavor in cognitive neuroscience. The growing interest in individual differences research has led to the development of new paradigms that may allow for novel insights into individual brain architecture. Naturalistic viewing (NV) is a promising tool for designing more ecologically valid functional magnetic resonance imaging (fMRI) studies, thus providing the opportunity to measure individual differences under beneficial circumstances (Vanderwal et al., 2017). In this regard, naturalistic paradigms represent a middle ground between task-based and resting-state fMRI. On the one hand, movies provide less artificial constraint than tasks (Finn et al., 2017) while still guiding the participants attention, which allows synchronizing brain states across participants, unlike resting-state fMRI. Due to advantageous participant engagement and compliance, movie fMRI appears promising for clinical applications (Eickhoff et al., 2020) or specific populations such as children (Vanderwal et al., 2019). However, while there is clearly a lot of potential movie fMRI, the effects of movie fMRI on individual variability of functional connectivity is yet unclear. This is especially the case in full-narrative movies (FNMs), representing a complete narrative arc rather than single movie scenes.

To utilize movie fMRI for studying individual differences with confidence, it is necessary to understand how movies influence variability of functional measures within and between participants. NV can induce high inter- and intra-subject correlations in activity time courses of various cortices (Hasson et al., 2004), which are dependent on features and content of the movie stimulus (Hasson et al., 2008; Lerner et al., 2011). These inter- and intra-subject correlations in brain activity are affected by the narrative coherence of a movie stimulus, as backward presentation of a movie decreases these correlations (Hasson et al., 2010). Moreover, movie stimuli can be edited to influence similarity as shown by higher inter-subject correlations in professionally produced movies than unedited, real-life movies (Hasson et al., 2010). A direct comparison between different movie stimuli and RS indicated that a complex movie with social interactions yielded higher inter-subject correlations compared to an abstract, nonverbal movie, which in turn led to higher inter-subject correlations than RS (Vanderwal et al., 2015).

According to Finn et al. (2017), a major argument why movie fMRI might be an excellent paradigm for studying individual differences is the assumed beneficial ratio of inter- to intra-subject correlation induced by movies: On the one hand movies represent a common cognitive reference frame for subjects' brain states, thus decreasing irrelevant inter-subject variability, while on the other hand retaining a subject's most identifying features. This is denoted by low intra-subject variability, that is, subjects being similar to themselves, for example, over the course of watching a movie or when watching the same movie in two separate sessions (Finn et al., 2017). Concordantly, a study by Vanderwal et al. (2017) showed that movies overall significantly decreased both inter- and intra-subject variability on a whole-brain level compared to RS, thus lending support to the idea that NV might preserve or even enhance individual differences in functional connectivity (Vanderwal et al., 2017). A recent study found that influences of NV on inter- and intra-subject similarity in NFC is dependent on the brain network and stimulus (Kröll et al., 2023), however, not all factors contributing to the effects of movies on inter- and intra-subject similarity of NFC have been investigated. For example, the impact of specific content, such as emotions portrayed in movies, is still unknown.

In contemporary neuroscience, functional networks consisting of distributed but interacting brain regions are often viewed as the foundation of cognition functions (Eickhoff & Grefkes, 2011), thus allowing for an interpretation of interactions between various brain regions with respect to specific cognitive domains. Given that movies are complex and multimodal stimuli that elicit widely distributed brain activity, a network perspective might help to untangle the effect of movies on specific cognitive functions. Specifically, studying the functional connectivity between the brain regions constituting these networks (i.e., network functional connectivity [NFC]) might grant insights into the effects of NV on brain function. While there are various methods to define functional networks (e.g., Fox et al., 2005; Pervaiz et al., 2020; Power et al., 2011; Schaefer et al., 2017; Smith et al., 2009; Yeo et al., 2011), one approach that instrumentalizes the body of existing knowledge about specific cognitive processes is the use of meta-analysis. Coordinate-based meta-analyses of neuroimaging data (e.g., activation likelihood estimation; Eickhoff et al., 2009) identify brain locations that are consistently activated during cognitive tasks across various studies. Converging results from many studies using different tasks to study the same cognitive function leads to a robust mapping of function-related brain coordinates (Eickhoff et al., 2012). In turn, reliably co-activated regions can be assumed to constitute a network that is engaged with the specific cognitive function (Fox et al., 2015). Various meta-analytical networks have been characterized, covering different psychological domains, and have been proven useful for gaining insight into the role of brain regions in a network perspective (Igelström & Graziano, 2017), robustly assessing the neural basis of cognitive functions (Binder & Desai, 2011; Etkin et al., 2015; Gross, 2015; Margulies et al., 2016) and therefore laying the ground for further experimental work (Morawetz et al., 2017). Studies using meta-analytical networks to predict personality scores (Nostro et al., 2018) or classify participants according to their mental health status and age (Pläschke et al., 2017) yielded better or at least similar results to using whole-brain connectivity (Nostro et al., 2018), while improving interpretability. Therefore, meta-analytical networks provide an excellent basis for studying individual variability in different cognitive systems.

With respect to movie fMRI, Vanderwal et al. (2017) showed that the effect of movies is differentially distributed across the brain, with lower inter- than intra-subject variability in unimodal regions and higher inter- than intra-subject variability in heteromodal regions. However, a concrete comparison of these variabilities on the level of NFC has rarely been done (but see Kröll et al., 2023).

Furthermore, it is yet unclear which features of a movie stimulus influence inter- and intra-subject similarity. On the one hand,

lower-level audiovisual features can be used to characterize movie stimuli (Cutting, 2016), on the other hand, conventional movies are hallmarked by their socio-affective content. Emotions are a main focus of most conventional movies reflecting social relationships and interactions in real life. Additionally, emotionally evocative narrative events were shown to be particularly good at synchronizing subjects (Chang et al., 2021).

Movies have been used to elicit basic positive and negative emotions (Gross & Levenson, 1995; Schaefer et al., 2010; Westermann et al., 1996), to investigate commonly co-occurring emotions (Gilman et al., 2017), and mixed emotions (Aaron et al., 2018; Kreibig et al., 2013; Kreibig et al., 2015; Kreibig & Gross, 2017; Samson et al., 2016). In movies with social content, the emotions portrayed by characters are important cues for eliciting emotions in viewers (Labs et al., 2015, Lettieri et al., 2019), making portrayed emotions an important stimulus feature in NV studies.

However, movies used for emotion induction are usually short clips to induce basic emotions (Gross & Levenson, 1995; Jenkins & Andrewes, 2012; Schaefer et al., 2010). Most of these clips are under 1 min in duration, with only some lasting a few minutes. With the advent of NV for individual differences and socio-affective research, studies started employing longer stimuli or multiple sessions (Alexander et al., 2017; Di Oleggio et al., 2017; Jääskeläinen et al., 2008; Kröll et al., 2023; Nguyen et al., 2017; Vanderwal et al., 2015; Vanderwal et al., 2017), but so far only very few employed complete or minimally shortened conventional movies (Hanke et al., 2014; Kauttonen et al., 2018). The duration of movie stimuli is a relevant aspect, as some cognitive and emotional processing only evolves over longer time frames (Hasson et al., 2010). Context is essential when understanding social situations and identifying or emphasizing with different characters. Arguably, more complex socio-affective processes can only be studied if the stimulus represents the full complexity of social relationships and emotions, such as in an FNM which spans the complete narrative arc of a conventional movie.

Thus, the advantage of showing FNMs is that emotions are presented embedded in context and progressing over time, which allows for simultaneously investigating multiple emotion components and how they develop and interact dynamically (Lettieri et al., 2022; Saarimäki, 2021; Sonkusare et al., 2019). Movies allow studying the association between emotion profiles, affectives states and associated brain states (Lettieri et al., 2019) as well as the effects of the movie stimuli on emotion-related brain networks across time (Nummenmaa et al., 2012). On top of that, the usage of FNMs allows the investigation of neural activation during sustained emotional arcs and the transitions between different emotional states. Therefore, an FNM might be used as a proxy for emotional real-life experiences, which facilitates the exploration of how the viewer's emotional engagement with the narrative modulates functional brain architecture. Exploring and understanding the potential of FNM for the study of individual differences is an extensive endeavor which needs to be approached from multiple angles: studying various movie stimuli, characterizing these stimuli in annotations, and investigating their effects on brain measures. We here take a first step to fill this wide gap by investigating inter- and intra-subject similarity in NFC over the course of an FNM. We used the publicly available studyforrest dataset which is unique in its length and overarching narrative, extended by an annotation of the emotions portrayed in the movie stimulus. This dataset contains fMRI acquisitions of subjects watching the full narrative arc of the popular movie "Forrest Gump."

In preliminary analyses, we first compared different segments of the movie stimulus with regard to their portrayed valence and arousal, evincing differences in emotional content. Using linear mixed models (LMMs), we then analyzed how different factors, such as the narrative of the movie and portrayed valence and arousal, affect inter- and intra-subject similarity in NFC across 14 meta-analytically defined networks.

We expected to see differences in inter- and intra-subject similarity that are network-dependent, change over the course of the FNM and are influenced by valence and arousal portrayed within the movie.

2 | METHODS

2.1 | Sample

This sample consisted of 15 native German-speaking participants (6 females, range 21–39 years) (Hanke et al., 2016). One subject, which we excluded, was an outlier in the intra-subject correlation analysis, leading to a sample size of 14 (6 females, age range 21–39 years. Please note that mean age cannot be reported, because only age ranges were reported for each participant). The Ethics committee of Otto-Von-Guericke University, Germany approved acquisition of the data in the "studyforrest" project. For a more detailed description of the sample, see Sengupta et al. (2016). The full dataset can be found under: https://github.com/psychoinformatics-de/studyforrest-data-phase2. A list of subjects can be found in Supplementary Table S1.

2.2 | MRI data and preprocessing

fMRI data acquisition took place in a single session which included a short break in the middle. To keep the stimulus at a length of 2 h, some scenes were cut. The movie stimulus represents an FNM, as only scenes that were less relevant to the plot were cut, thus preserving the overarching story. For the purpose of data acquisition, the movie stimulus was separated into eight segments of approximately 15 min each, taking scene boundaries into consideration. This lead to an unequal number of volumes acquired per segment, which were 451, 441, 438, 488, 462, 439, 542, and 338 for segments 1–8, respectively (see Hanke et al., 2014 for details and code on movie segment creation). The movie segments were shown in chronological order. For each segment, T2*-weighted echo-planar images (gradient-echo, 2 s repetition time [TR], 30 ms echo time, 90° flip angle, 1943 Hz/Px bandwidth, parallel acquisition with sensitivity encoding [SENSE] reduction factor 2, 35 axial slices, 3.0 mm slice thickness, 80 × 80 voxels [3.0 × 3.0 mm] in-plane resolution, 240 × 240 mm field-of-

view, anterior-to-posterior phase encoding direction in ascending order, 10% inter-slice gap, whole-brain coverage) were acquired using a whole-body 3 Tesla Philips Achieva dStream MRI scanner equipped with a 32 channel head coil.

All downloaded data were minimally preprocessed as described in Hanke et al. (2016). In short, preprocessing steps included defacing, motion correction, reslicing and data interpolation using in-house codes that utilize the FSL toolkit. All codes are openly available under: https://github.com/psychoinformatics-de/studyforrest-data-aligned/tree/master/code. For precise information about observed motion and data quality analyses, see Hanke et al. (2016). For this study, the native fMRI data were brought into MNI space using FSL's applywarp function for subsequent NFC extraction.

2.3 | Valence and arousal measures

To characterize the movie segments with regard to the portrayed valence and arousal, we used the openly available data from Labs et al. (2015). This dataset contains annotations of portrayed emotions in the "Forrest Gump" movie stimulus used in the "studyforrest" dataset.

A group of observers ($n = 9$, German female university students) were asked to evaluate scenes of the movie in terms of valence ("positive" or "negative") and arousal ("high" or "low") portrayed by the movie characters. All scenes were presented in random order to allow observers to focus on current indicators of portrayed emotions without being influenced by, for example, the conveyed mood of the movie plot. To evaluate the consistency of evaluations between observers, Labs and colleagues calculated the inter-observer agreement (IOA). The IOA value describes the portion of observers indicating the presence of a specific attribute in a scene (Labs et al., 2015). As arousal and valence were measured on a bipolar scale ("positive" of "negative" valende, "low" or "high" arousal present), the timeseries of these attributes were calculated as the difference between the IOA timeseries of both expressions. That is, the IOA timeseries of arousal was calculated by subtracting the IOA timeseries of low arousal segments from the IOA timeseries of high arousal segments (Labs et al., 2015). The IOA is expressed as a value between 1 and −1, with "1" indicating perfect observer agreement regarding high arousal (or positive valence, respectively) and "−1" indicating perfect observer agreement regarding low arousal (or negative valence, respectively). IOAs were reported as a time series of the movie in correct order downsampled to 2 s, corresponding to the sampling rate of the fMRI data. We used code published by Lettieri et al., 2019 (https://osf.io/tzpdf/) to divide the IOA time series according to 8 movie segments for subsequent analyses.

2.4 | Inter- and intra-subject similarity in functional networks

To investigate effects on a network level, we used 14 networks defined as sets of peak coordinates in different meta-analyses. These included the autobiographical memory (AM) network (Spreng & Grady, 2010), cognitive attention control (CogAC) network (Cieslik et al., 2015), extended multiple demand network (eMDN) (Camilleri et al., 2018), emotional scene and face processing (EmoSF) network (Sabatinelli et al., 2011), empathy network (Bzdok et al., 2012), theory of mind (ToM) network (Bzdok et al., 2012), emotion regulation (ER) network (Buhle et al., 2014), extended socio-affective default network (eSAD) (Amft et al., 2015), mirror neuron system (MNS) network (Caspers et al., 2010), motor network (Witt et al., 2008), reward (Rew) network (Liu et al., 2011), semantic memory (SM) network (Binder et al., 2009), vigilant attention (VigAtt) network (Langner & Eickhoff, 2013), and the working memory (WM) network (Rottschy et al., 2012). A more detailed description of these networks are reported in the supplements (Supplementary Material S2). For each meta-analytical network, nodes were created by placing 6 mm spheres around the peak coordinates (see Supplementary Material S3 for an overview of the peak coordinates and S4 for a figure of all nodes of all networks). The functional connectome of a given network was created using in-house MATLAB R2017a (The Mathworks Inc., 2017) code which computes the pairwise Pearson correlation between all nodes for each segment and each participant. This resulted in 1680 functional network connectomes (15 participants × 8 segments × 14 networks) saved as N-by-N matrices with N being the number of nodes.

To keep in line with previous studies (Finn et al., 2017; Nastase et al., 2019; Vanderwal et al., 2015; Vanderwal et al., 2017), we operationalized the inter- and intra-subject similarity as the Pearson correlation coefficients between functional connectomes within and between subjects. Inter- and intra-subject similarity were computed per network, segment and subject as depicted in Figure 1. All computations are based on the unique connections between nodes (i.e., the lower triangle of the NFC matrix) and exclude all autocorrelations. For inter-subject similarity, we first computed the correlations between the NFC of one subject and all other subjects. After Fisher Z-transformation of the correlation coefficient, they were averaged and re-transformed, resulting in one value representing inter-subject similarity for the respective subject in the given segment and network. For calculating intra-subject similarity of a given subject and segment, we computed the correlations between NFCs of this segment and every other segment of the subject. The correlation values were Fisher's z-transformed, averaged, and reverted to *r*-values, resulting in one value representing intra-subject similarity for the respective subject in the given segment and network. Both inter- and intra-subject similarity were calculated based on Pearson correlation coefficients.

2.5 | Statistical analyses

To investigate whether portrayed emotions are different across movie segments, we conducted a one-way ANOVA for each measure. Here, IOA values were used as independent variables with the movie segments as fixed factors. Post hoc *t* tests are reported with Bonferroni-adjusted *p* values.

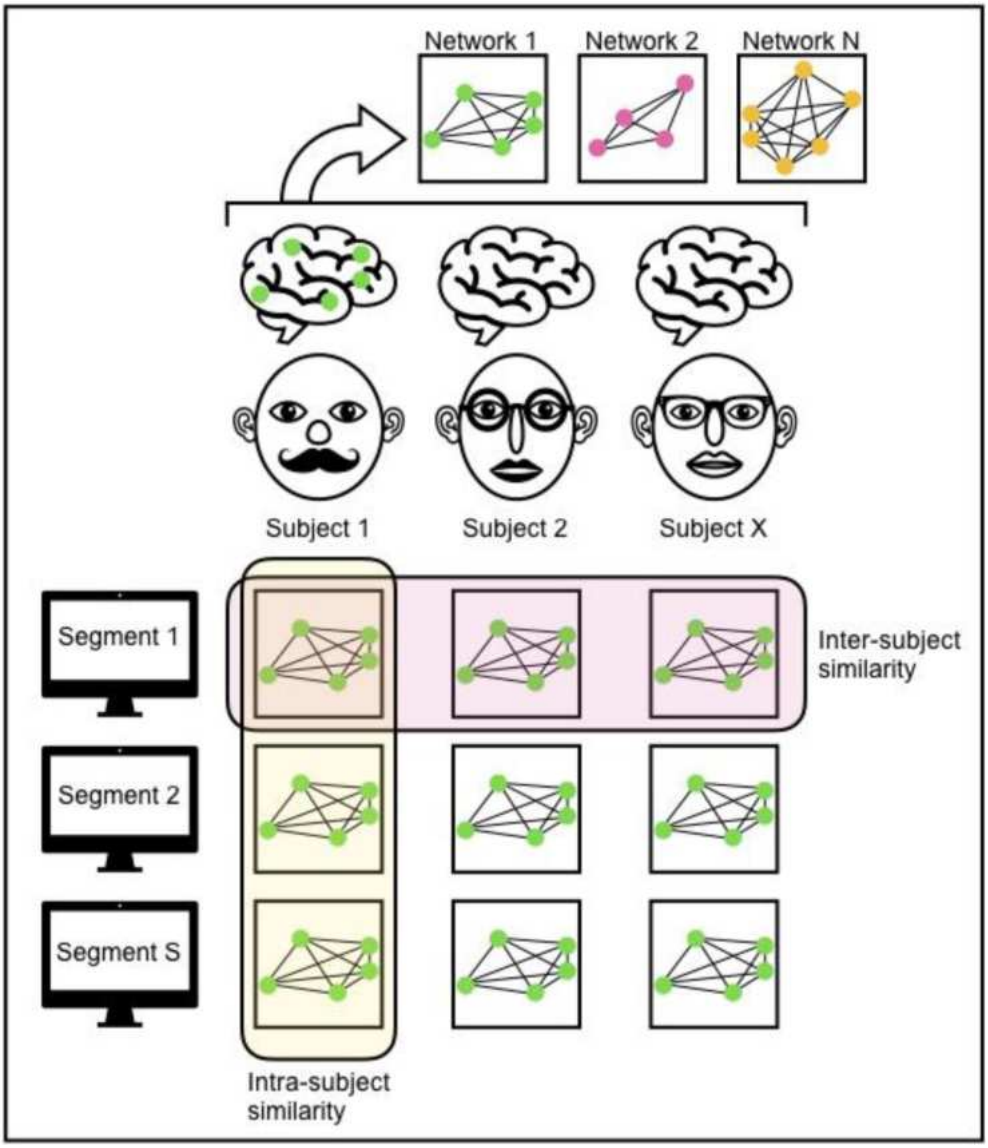


FIGURE 1 Calculation of inter- and intra-subject similarity. For each subject, functional connectomes were computed for all 14 networks in each of the 8 movie segments. Inter-subject similarity is assessed by calculating the average correlations between subjects within the same movie segments. Intra-subject similarity is assessed by calculating the average correlation between movie segments within the same subject.

To test whether inter- or intra-subject similarity differ across networks depending on movie segments and portrayed emotions, we applied LMMs using the statsmodels python package (https://www.statsmodels.org/stable/mixed_linear.html). Specifically, we created different random intercept models by choosing network, movie segment, arousal and valence as possible fixed effects, subject identity as a random effect, and inter- or intra-subject similarity as the dependent variable. We chose subject identity as a random effect, because participants are the sampling unit of interest and contribute repeatedly to the NFC measures across all movie segments. We model individual differences by assuming different random intercepts for each subject, but no individual random slopes, because a simpler model structure was warranted by our data. Network was chosen as a fixed effect to test which networks are associated with changes in inter- or intra-subject similarity induced by NV. It was included as a categorical factor with 14 levels. Movie segment was chosen as a fixed effect to test for an effect of the length and complexity of an FNM. Portrayed valence and arousal were chosen as fixed effects to represent the emotional content of the FNM, testing if emotions portrayed in a movie affect inter- or intra-subject similarity in NFC. Models were generated using maximum likelihood to include all possible models, that is, each unique combination of one to four fixed effects and their respective interactions, resulting in 2128 models that were compared each for inter- and intra-subject similarity. The model best fitting our data were selected using Bayesian information criterion (BIC, Schwarz, 1978) and used to calculate the parameter estimates for each effect. To test whether a specific network had a significant influence on inter- or intra-subject similarity, we created a "mean network" representing the mean inter- or intra-subject similarity values across all networks that we used as a reference category to compare all other networks against. *p*-Values were obtained using Wald tests of the best models.

3 | RESULTS

Emotions and an overarching narrative are hallmark features of conventional Hollywood movies, which are frequently employed in NV

research because of their engaging and complex nature. However, most NV studies use only shorter clips from these movies, essentially excluding effects of the ongoing narrative. Therefore, is it not yet clear how these features might impact inter- and intra-subject similarity in NFC in an FNM. Here, we investigated portrayed valence and arousal across an FNM and how these factors contribute to explaining inter- and intra-subject similarity in NFC in 14 networks.

3.1 | Movie segments and portrayed emotions

We used a previously reported description (Labs et al., 2015) of portrayed valence and arousal for comparisons between the emotional content of different movie segments. Our results showed that movie segments differed in the direction (i.e., positive/negative valence; high/low arousal) and the extent of agreement between observers concerning these measures (Figure 2). Figure 2 shows average IOA values of each movie segment and reveals large differences in the evaluation of valence and arousal across movie segments. For segments 1, 6, 7, and 8 IOA values indicate consistency in portrayed positive valence, while the segments 4 and 5 portrayed negative valence. Segment 2 and 3 showed little consistency in the evaluation of portrayed valence, as IOA values are close to zero. Concordantly, the ANOVA on the valence IOA values resulted in a significant main effect of segment ($F(7,3534) = 45.879$, $p < .001$), and Bonferroni-corrected post hoc testing revealed significant differences between the consecutive segments 1 and 2 ($t = 3.378$, $p = .021$); 3 and 4 ($t = 7.236$, $p < .001$); 4 and 5 ($t = -3.131$, $p = .049$); 5 and 6 ($t = -8.519$, $p < .001$); and 7 and 8 ($t = 3.552$, $p = .011$). Segment 4 had the strongest agreement on negative valence between observers, while segment 7 showed the strongest agreement on positive valence between observers. Figure 2 further shows that segments 2 and 4 portrayed high arousal, while the other segments portrayed low arousal. The ANOVA on arousal IOA values showed a significant main effect of segment as well ($F(7, 3534) = 15.479$, $p < .001$). Bonferroni-corrected post hoc testing revealed significant differences between consecutive segments 1 and 2 ($t = -13.448$, $p < .001$); 2 and 3 ($t = 10.397$, $p < .001$); 3 and 4 ($t = -14.628$, $p < .001$); and 4 and 5 ($t = 10.617$, $p < .001$).

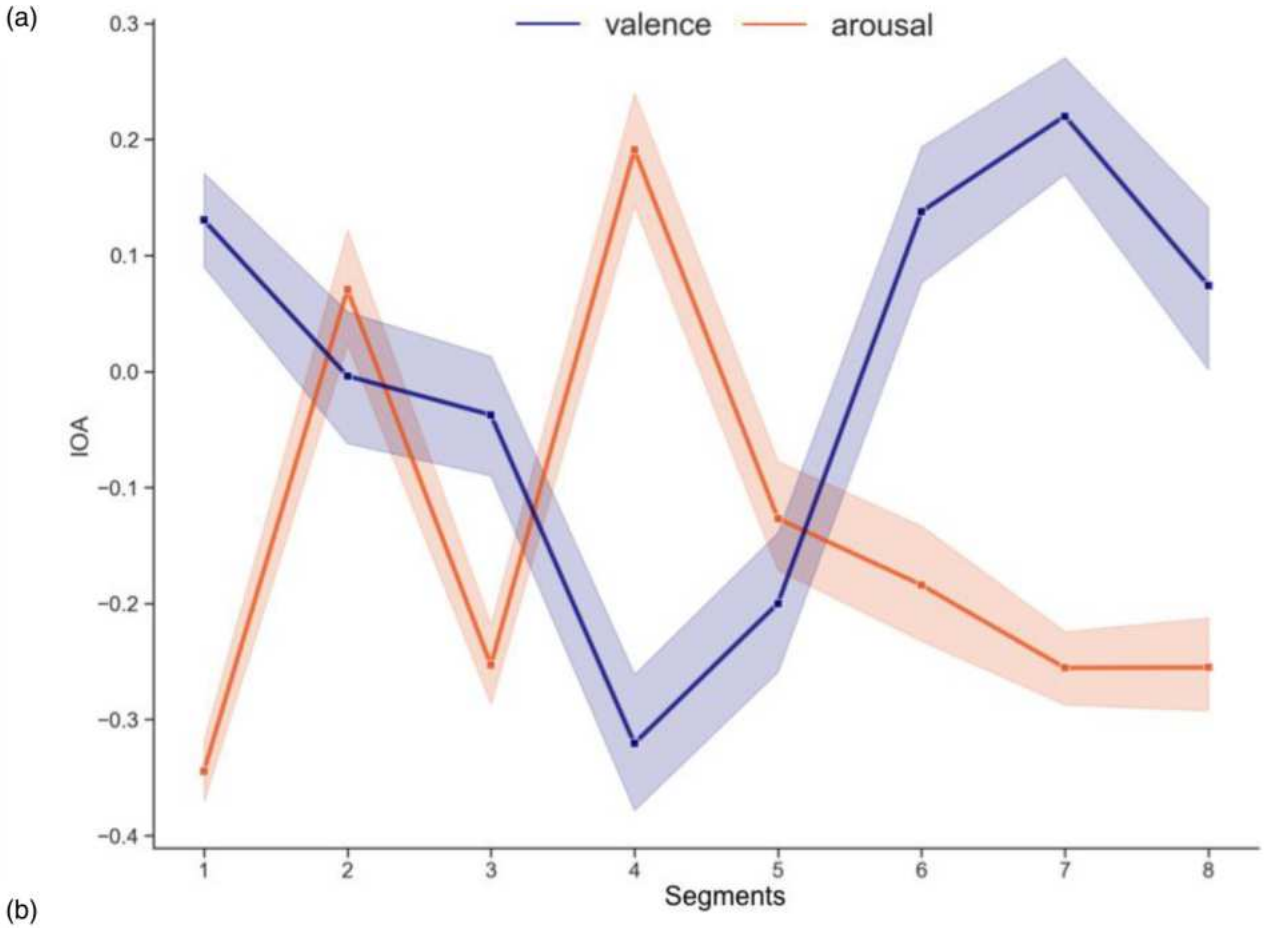
Given how much portrayed emotions and the narrative are intertwined, our results are best interpreted in the light of the content of the movie segments. Segment 1 spans the introduction of Forrest Gump and scenes from his childhood, containing both positive (caring mother, close friendship with neighbor girl Jenny) and negative (walking impairments, bullying) elements. Segment 2 was marked by low IOA in both valence and arousal, showing less agreement between observers on the portrayed emotions in this segment. During this segment, the movie shows Forrest's highschool and college time, addressing athletic successes and first dating experiences. Low IOA values continue in the valence dimension in segment 3, whereas observers agreed more strongly on low arousal being portrayed here. Here, the movie shows Forrest joining the army, reuniting with Jenny in a nightclub where she works as a dancer, and being deployed in the Vietnam war. Segment 4 prominently features a different pattern than any other segment: observers agreed that movie characters displayed high arousal and low valence during this segment. This can likely be attributed to the war scenes involving an ambush in Vietnam causing Forrest's best friend's death, and following scenes in a military hospital, although the segment also contains Forrest receiving the Medal of Honor and speaking at an anti-war rally in front of the Pentagon. Segment 5 is marked by lower IOA values indicating some negative valence and low arousal, featuring the Black Panther movement, Forrest's ping pong career and reunions with friends Jenny and Lt. Dan. The last three segments again display a pattern of higher agreement between observers on positive valence and low arousal, when the movie spans Forrest's successful shrimp fishing business, two episodes of living happily with Jenny, a three-year cross-country marathon, Forrest meeting his son, Jenny's death and the ending of the movie.

3.2 | Inter- and intra-subject similarity in NFC

Inter- and intra-subject correlations were calculated for every network on the level of single segments, that is, the different segments of the movie. Figure 3 summarizes the results across all networks and segments based on Pearson correlation coefficients (Figure 3a), and shows results of the LMM analyses on inter- and intra-subject (Figure 3b) similarity (Figure 3c). We found that inter- and intra-subject similarity both fluctuate across time for all networks. We will further analyze the statistical significance in the following sections.

3.2.1 | Inter-subject similarity

The best model that best fitted on inter-subject similarity as selected using BIC consists of the random factor subject identity, the fixed factors network, movie segment and valence, and the interaction between the fixed factors movie segment and valence. All parameter estimates and p values can be seen in Figure 3b. The intercept for inter-subject similarity is 0.245, representing the average inter-subject correlation value. Of all 14 networks, the AM, ER, EmoSF, Empathy, Rew, SM, VigAtt, WM, and eMDN networks differed significantly from the "mean network" reference category representing the mean inter-subject similarity across all networks. The AM, ER, and SM networks show negative coefficients, indicating that inter-subject similarity is lower in these networks than on average. The EmoSF, Empath, Rew, VigAtt, WM, and eMDN networks were associated with higher inter-subject similarity than average. Movie segment, valence and their interaction effect reached significance as well. While movie segment and the movie segment-valence interaction were associated with higher inter-subject similarity, valence was associated with lower inter-subject similarity. The estimated coefficient for subject identity was 0.001, indicating a low effect of subject identity on inter-subject similarity and small differences between subjects.



(b)

Valence	S1	S2	S3	S4	S5	S6	S7
S1							
S2	3.378						
S3	4.212	0.853					
S4	11.630	8.107	7.236				
S5	8.401	4.948	4.095	-3.131			
S6	-0.185	-3.535	-4.363	-11.722	-8.518		
S7	-2.371	-5.879	-6.745	-14.587	-11.162	-2.157	
S8	1.319	-1.822	-2.599	-9.410	-6.455	1.482	3.552

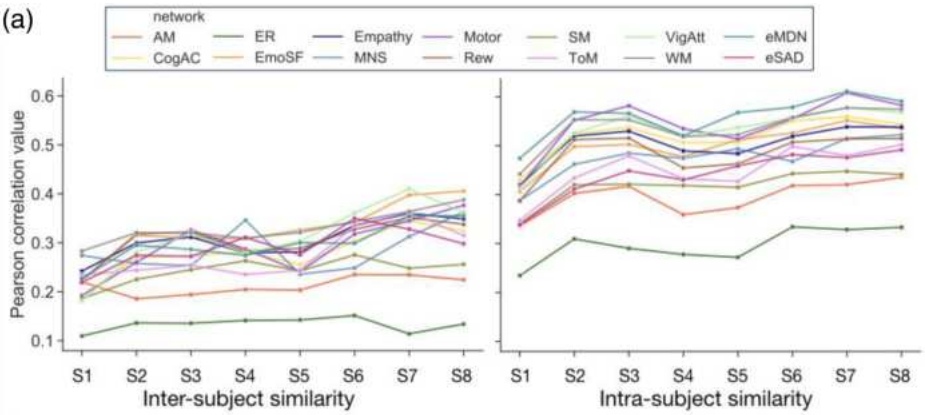
Arousal	S1	S2	S3	S4	S5	S6	S7
S1							
S2	-13.448						
S3	-2.957	10.397					
S4	-17.799	-3.976	-14.628				
S5	-7.139	6.424	-4.106	10.617			
S6	-5.196	8.180	-2.218	12.359	1.860		
S7	-3.034	11.026	0.081	15.526	4.410	2.416	
S8	-2.704	9.788	0.060	13.705	3.897	2.136	-0.012

FIGURE 2 Results of the ANOVA on valence and arousal inter-observer agreement (IOA) in each movie segment. (a) Valence and arousal IOA across movie segments (portrayed valence: purple and arousal: orange). Positive IOA values indicate that observers agreed on the portrayal of positive valence and high arousal, while negative IOA values indicate that observers agreed on the portrayal of negative valence and low arousal. The amount of deviation from zero in IOA values corresponds to the strength of agreement between observers. For each movie segment, the IOA values are averaged across the entire segment. (b) Post hoc results of the ANOVA on valence (left) and arousal (right). Bonferroni-corrected significance levels are represented by colors: Orange signifies p values < .001, yellow marks p values < .05 and white marks no significance. S1–S8 = segments 1–8. Direction of the t tests is column minus row element.

3.2.2 | Intra-subject similarity

The best model that best fitted on intra-subject similarity as selected using BIC consists of the random factor subject identity, the fixed factors network, movie segment, valence and arousal, and the interactions between fixed factors movie segment and valence and between movie segment, arousal and valence. All parameter estimates and p values can be seen in Figure 3c. The intercept for intra-subject similarity is 0.473, representing the average intra-subject correlation value. The AM, CogAC, ER, EmoSF, Empathy, Motor, SM, ToM, VigAtt, WM, eMDN, and eSAD network differed significantly from the reference category representing the mean intra-subject similarity across all networks. The AM, ER, SM, ToM, and eSAD networks were associated with lower intra-subject similarity, whereas the CogAC,

(a)



(b)

INTER	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	0.245	0.011	23.028	0.000	0.224	0.266
network[T._AM]	-0.066	0.008	-8.641	0.000	-0.081	-0.051
network[T._CogAC]	0.002	0.008	0.238	0.812	-0.013	0.017
network[T._ER]	-0.146	0.008	-19.073	0.000	-0.161	-0.131
network[T._EmoSF]	0.050	0.008	6.520	0.000	0.035	0.065
network[T._Empathy]	0.029	0.008	3.743	0.000	0.014	0.044
network[T._MNS]	0.008	0.008	1.002	0.316	-0.007	0.023
network[T._Motor]	0.015	0.008	1.911	0.056	0.000	0.030
network[T._Rew]	0.028	0.008	3.634	0.000	0.013	0.043
network[T._SM]	-0.036	0.008	-4.721	0.000	-0.051	-0.021
network[T._ToM]	-0.001	0.008	-0.102	0.919	-0.016	0.014
network[T._VigAtt]	0.033	0.008	4.346	0.000	0.018	0.048
network[T._WM]	0.053	0.008	6.880	0.000	0.038	0.068
network[T._eMDN]	0.020	0.008	2.655	0.008	0.005	0.035
network[T._eSAD]	0.012	0.008	1.609	0.108	-0.003	0.027
movie segment	0.006	0.001	6.220	0.000	0.004	0.008
valence	-0.311	0.038	-8.082	0.000	-0.386	-0.235
movie segment:valence	0.073	0.008	8.833	0.000	0.057	0.089
subject Var	0.001	0.007				

(c)

INTRA	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	0.473	0.029	16.279	0.000	0.416	0.529
network[T._AM]	-0.080	0.011	-7.091	0.000	-0.103	-0.058
network[T._CogAC]	0.044	0.011	3.874	0.000	0.022	0.066
network[T._ER]	-0.178	0.011	-15.707	0.000	-0.200	-0.156
network[T._EmoSF]	0.025	0.011	2.242	0.025	0.003	0.048
network[T._Empathy]	0.029	0.011	2.548	0.011	0.007	0.051
network[T._MNS]	0.000	0.011	0.012	0.990	-0.022	0.022
network[T._Motor]	0.067	0.011	5.927	0.000	0.045	0.090
network[T._Rew]	0.008	0.011	0.693	0.489	-0.014	0.030
network[T._SM]	-0.057	0.011	-5.056	0.000	-0.080	-0.035
network[T._ToM]	-0.026	0.011	-2.254	0.024	-0.048	-0.003
network[T._VigAtt]	0.056	0.011	4.958	0.000	0.034	0.079
network[T._WM]	0.062	0.011	5.427	0.000	0.039	0.084
network[T._eMDN]	0.084	0.011	7.397	0.000	0.062	0.106
network[T._eSAD]	-0.034	0.011	-2.971	0.003	-0.056	-0.011
movie segment	0.002	0.002	0.981	0.327	-0.002	0.005
arousal	0.015	0.023	0.630	0.529	-0.031	0.061
valence	-0.771	0.090	-8.601	0.000	-0.946	-0.595
movie segment:valence	0.180	0.018	10.151	0.000	0.145	0.215
movie segment:arousal:valence	0.172	0.021	8.253	0.000	0.131	0.213
subject Var	0.010	0.046				

not sign.
p < .05
p < .001

AM = autobiographical memory, CogAC = cognitive attention control, ER = emotion regulation, EmoSF = emotional scene and face processing, MNS = mirror neuron system, Rew = reward, SM = semantic memory, ToM = theory of mind, VigAtt = vigilant attention, WM = working memory eMDN = extended multiple demand network, eSAD = extended socio-affective default mode network, S1-S8 = segments 1-8

FIGURE 3 Results of the LMM showing how the fixed effect network, movie segment, arousal and valence and the random effect subject identity contribute to inter- and intra-subject similarity in NFC. (a) Inter- and intra-subject correlation of each movie segment averaged in each network, based on Pearson correlation coefficients. The x axis depicts different movie segments. The y axis represents the averaged similarity of the functional connectivity matrix derived from one subject compared to all other subjects within each network. (b) Results of the LMM on inter-subject similarity. (c) Results of the LMM on intra-subject similarity.

EmoSF, Empathy, Motor, VigAtt, WM, and eMDN networks were associated with higher intra-subject similarity than average. While movie segment and arousal did not reach significance, valence, the movie segment-valence interaction and the movie segment-arousal-valence interaction did. Valence was associated with lower intra-subject similarity, while the movie segment-valence and movie

segment-arousal-valence interaction were associated with higher intra-subject similarity. The estimated coefficient for subject identity was 0.01, indicating a low effect of subject identity on intra-subject similarity and small differences between subjects.

4 | DISCUSSION

The results of this study present a first step in understanding the potential of FNM for the study of individual differences, in particular with respect to the emotional arcs evolving over the course of the movie and transitions between different emotional states. By analyzing a publicly available dataset that contains fMRI data spanning an FNM, we investigated changes in inter- and intra-subject similarity over multiple, consecutive movie segments. Inter- and intra-subject similarity were best explained when accounting for network, movie segment, valence and a movie segment by valence interaction. Additionally, arousal played a role in explaining intra-subject similarity by interacting with movie segment and valence. The effect of the movie stimulus on changes in inter- and intra-subject similarity was network dependent. Comparing portrayed valence and arousal across movie segments showed that both varied across the segments, indicating differences in emotional content that we could relate to the content of the different movie segments.

4.1 | Portrayed valence and arousal

Emotions are an important feature of movie stimuli. Shorter movies have been used to study emotion processing (Carvalho et al., 2012; Schaefer et al., 2010; Westermann et al., 1996) and longer movies might allow studying emotions across a larger timescale. Emotions are a major factor in narration (Aldama, 2015; Cutting, 2016), change over time, and dynamically interact with social context (Redcay & Moraczewski, 2020). Therefore, FNMs have clear advantages for studying emotions in a naturalistic setting. Additionally, emotions portrayed in movies affect inter-subject synchronization (Dziura et al., 2021) and inter-subject alignment of brain states (Chang et al., 2021), which makes them a relevant factor for studying individual differences using NV.

Here, we used a previously reported description (Labs et al., 2015) of portrayed valence and arousal for comparisons between the emotional content of different movie segments. Our results showed that movie segments differed in the direction (i.e., positive/negative valence; high/low arousal) and the extent of agreement between observers concerning these measures (Figure 2). In our results a pattern emerged in which segments that were marked by high concordance in positive valence also showed good agreement in low arousal evaluations (segments 1, 6, 7, 8), whereas the reversed pattern was observable in segment 4. This indicates a potential negative relationship between valence and arousal as depicted in our movie stimuli. Valence and arousal are the bipolar dimensions in circumplex models of affect (Yik et al., 1999), and the relationship between valence and arousal seems to be highly individual and related to personality and culture (Kuppens et al., 2017).

Overall, the pattern of results implies that the segments of the chosen movie stimulus differed in emotional content, which makes it valuable for inducing variability in functional networks associated with socio-emotional processing. Specifically, the Forrest Gump movie features a broad range of themes (love, friendship, politics, fate); settings (varying historical events, places, times and roles of the protagonist); and situations portraying a wide spectrum of emotions in different contexts. Our results are thus in line with studies showing that movies can elicit complex and mixed states of emotions (Carvalho et al., 2012; Schaefer et al., 2010). In particular, the Forrest Gump movie stimulus has been shown to induce distinct affective states throughout the movie, which was used to map the topographic organization of these states (Lettieri et al., 2019). Hence, in accordance with the proposal by Finn et al. (2017), the chosen movie can evoke brain states in a meaningful manner, and thus represents a fitting stimulus for studying variability in and between subjects over time.

We investigated portrayed valence and arousal as important emotional features of the movie stimulus. Critically, Labs et al. (2015) created an annotation of the movie stimulus content, not an annotation of the viewer's emotional experiences. In order to characterize the portrayed emotions as a relatively lower level feature, observers rated all movie scenes in randomized order to prevent "carry-over" effects from the context the scenes appear within and the current mood of the movie (Labs et al., 2015). This annotation therefore offers descriptive information about the movie stimulus rather than assessing the full emotional complexity of the movie and its effects on the viewer. The characterization of emotion cues in single scenes offers the benefit of relating these cues to other features of the movie scenes (e.g., lighting, audio features) in future studies.

4.2 | Inter-subject similarity

Across networks, inter-subject correlations increased over the course of the movie, indicating a general tendency of subjects' NFC to become more similar (Figure 3a).

By using LMM, we found several factors contributing to changes in inter-subject similarity, including network, movie segment, valence and interaction of movie content and valence. Specifically, the model that best explains inter-subject similarity comprises the fixed effects network, movie segment, valence and a movie segment by valence interaction, with subject identity as a random effect.

Looking more closely at the networks, we see that some networks, such as the CogAC, MNS, Motor, ToM, and eSAD, do not contribute significantly to changes in inter-subject similarity. This might indicate that these networks are not sensitive to the effects of an FNM and the emotions portrayed within. For those networks that are significantly modulated by movie content, we observed large variations across networks in inter-subject similarity. The AM, ER, and SM networks are associated with lower inter-subject similarity, which can be seen in lower inter-subject correlation values (Figure 3a) and negative model coefficients (Figure 3b). This indicates that ER and long term memory processes are most sensitive to an FNM. This might reflect the stimulus containing a highly emotional narrative and many

references to real world events and history. Contrarily, the EmoSF, Empathy, Rew, VigAtt, WM, and eMDN networks are associated with higher inter-subject similarity. Across all networks, the coefficients exhibit a wide range in values, with the ER network showing the highest absolute coefficient, indicating the greatest effect on inter-subject similarity. Movie segment had a small negative effect on inter-subject-similarity, indicating that inter-subject similarity increases over the course of an FNM, which is also reflected in a slight increase in inter-subject correlation values (Figure 3a). Previous research has shown high inter-subject variability in response to professionally produced and conventional movies that were much shorter (<20 min) than an FNM (Hasson et al., 2010; Vanderwal et al., 2015). It is likely that the change toward more similarity in NFC over the course of the movie results from the shared experience, which is created to evoke certain reactions and feelings in the audience. Indeed, viewers' emotional and cognitive states can be affected and synchronized through director's decisions, such as the camera settings, light, performance of actors, scripts and dialog, and more (Baranowski & Hecht, 2017; Tarvainen et al., 2015). Studying viewers' emotions while watching the identical stimulus used here, Lettieri et al. (2019) showed that ratings of basic emotions were consistent across viewers, indicating an overall highly similar emotional experience induced by the movie. Emphasizing the relevance of affective states in movie fMRI, previous studies showed higher alignment of brain states between subjects during highly affective events in a TV show (Chang et al., 2021) and more synchronization of amygdala activity between subjects during positive events in a "shared watching" condition (Dziura et al., 2021).

In our study, valence was associated with lower inter-subject similarity. A study by Nummenmaa et al. (2012) studied the relationship between perceived valence and arousal and inter-subject synchronization of brain activity during movie watching. They found that more negative valence was associated with increased inter-subject synchronization in an emotion-processing network and the default-mode network, while high arousal was associated with increased inter-subject synchronization in somatosensory cortices, and visual and dorsal attention networks (Nummenmaa et al., 2012). This is in line with the pattern of positive valence being associated with lower similarity in our results.

However, the movie segment by valence interaction has a positive coefficient, indicating that positive valence is associated with higher inter-subject similarity across the course of an FNM. This might represent the effects of a conventional Hollywood movie orchestrating similarity in viewers' experience by using positive portrayed emotions.

The random factor subject identity had a very small negative effect on inter-subject similarity, indicating that there were no great differences between subjects.

4.3 | Intra-subject similarity

Our results show that intra-subject similarity increases over the course of an FNM across networks (see Figure 3a).

When selecting the best model in our LMM analysis to explain intra-subject similarity, network, movie segment, arousal, and valence emerged as relevant fixed effects. Additionally, the model included a movie segment by valence and a movie segment by arousal by valence interaction. Again, subject identity was included as a random effect.

Of all networks, only the MNS and Rew networks did not affect intra-subject similarity significantly. The AM, ER, SM, ToM, and eSAD networks were associated with decreased intra-subject similarity, while the CogAC, EmoSF, Empathy, Motor, VigAtt, WM, and eMDN networks were associated with increased intra-subject similarity. Similar to the results on inter-subject similarity, ER and long-term memory were most sensitive to the effects of an FNM, showing the lowest intra-subject similarity across movie segments. Additionally, networks processing self- and other-related social cognition showed low intra-subject similarity, indicating that an FNM might tax introspection and relating to others in a way that varies along the narrative.

Similar to the pattern of results seen in inter-subject similarity, valence was associated with lower intra-subject similarity while the movie segment by valence interaction was associated with higher intra-subject similarity. Additionally, the three-way interaction between movie segment, valence and arousal was associated with higher intra-subject similarity. This might indicate that positive valence is generally associated with lower intra-subject similarity, although the progression of movie segments and higher portrayed arousal increase intra-subject similarity. While arousal did not significantly influence inter-subject similarity, it interacts with movie segment and valence when influencing intra-subject similarity. This might indicate that arousal is a more relevant factor when investigating similarity within subjects and might prompt future comparisons on the effects of movies with different levels of arousal on single subjects. Arousal seems to be influenced by various stylistic features of a movie and can be further differentiated into subdimensions such as energetic and tense mood (Tarvainen et al., 2015).

The random factor subject identity had a very small positive effect on intra-subject similarity, indicating that there were no great differences between subjects.

4.4 | Limitations

Our study is one of the first that has used an FNM in the study of individual differences in brain organization and present results offers important insights into inter- and intra-subject similarity in NFC across a 2 h acquisition period. While these results are preliminary in the sense that they can not necessarily be generalized to other movies they clearly motivate the use of FNMs over the commonly used shorter movie segments. Our results indicate that the content of a movie is a relevant factor in NV, but it is not yet certain how different content or features of a movie relate to inter- and intra-subject similarity. Our study of one FNM and its annotation of portrayed valence and arousal is an important first step in quantifying this relationship. To generalize our results to other movies, brain measures and samples, future research needs to expand information on available NV

datasets (e.g., by creating more annotations), so that content and effects on NFC can be investigated across different datasets. It is necessary to find a good match between movies and their annotated features, methodology and research question (Eickhoff et al., 2020; Grall & Finn, 2022; Saarimäki, 2021).

Our study comes at the cost of investigating only the effects of a single movie. Comparisons with an equally long resting state acquisition or a movie stimulus without a narrative would have given stronger evidence for the effect of FNMs. However, there were no such scans available in this dataset. Analyses of additional FNMs might expand the insights gained into the effects of different narratives. The choice of a conventional Hollywood movie might have led to higher inter-subject similarity (Baranowski & Hecht, 2017; Chang et al., 2021; Hasson et al., 2010; Tarvainen et al., 2015; Vanderwal et al., 2015), while more ambiguous or emotionally and socially equivocal movies could enhance inter-individual differences to a greater degree.Familiarity with a movie stimulus has been discussed as a potential factor for driving individual differences. An effect of repeated movie watching in functional connectivity on the network level has been shown before (Andric et al., 2016). However, such effects can be assumed to be low in our sample. All participants were familiar with the narrative of the movie, and only one participant reported to never have seen the movie (Hanke et al., 2014).

Given the unusual length of data acquisition, effects of the MRI measurement might have influenced the participants' focus on and perception of the movie stimulus. For example, participants might have needed some time to familiarize themselves with the MRI scanner. However, as all participants had already participated in previous MRI measurements of the studyforrest project (Hanke et al., 2016; Sengupta et al., 2016), high familiarity to MRI scanning and all related procedures was present in this sample. The length of acquisition might also have affected the participants' attention. Previous studies indicate that movie watching is very engaging and might decrease drowsiness and sleep in the scanner (Eickhoff et al., 2020), but attention might still have been impacted over such a long duration. NV paradigms are designed to include minimal participant instructions so as not to influence participants' perception of the stimuli or add task demands not directly related to movie watching. In future studies, post hoc questionnaires might be useful to estimate attention fluctuations, distractions, drowsiness and other potential confounds that might have occurred during data acquisition.

The analyses of this study focused on the approximately 15-min segments the data were acquired in, splitting the movie into eight segments. Time windows for analysis of NV data have varied in the literature and optimal time windows and scan lengths are still debatable. Uri Hasson's work on temporal receptive windows focuses on window sizes on the level of seconds (e.g., time windows of ∼4 ["short"], ∼12 ["intermediate"], and ∼36 s ["long"]) (Hasson et al., 2010). Based on NV data, single subject identification accuracy was positively impacted by longer scan durations (Vanderwal et al., 2017; scan duration with highest accuracy ranged from ∼4.5 to ∼7 min depending on movie stimulus) and movies of ∼2.5 min length can be sufficient for behavioral prediction (Finn & Bandettini, 2021). Efforts for providing normative data during movie watching have been recommended to use minimally 10 min and optimally at least 25 min duration per movie (Eickhoff et al., 2020). Irrespective of NV, reliability of functional connectivity measures increases with time, with indications that less than 10 min of RS scan duration may not capture functional connectivity features reliably (Laumann et al., 2015, 9–27 min durations; Noble et al., 2017, 5–25 min durations). These examples show that optimal scan duration may depend strongly on the research question at hand, with advantages coming from longer durations. In our study, employing an FNM with the focus on a continuously unfolding and dynamic narrative might speak for longer scan durations to capture the effects of these "longer term" story dynamics.

While the long fMRI acquisition spanning an FNM is a great advantage to our study, it comes with the disadvantage of a small sample size. Replication in other datasets is an important next step, although this specific dataset currently remains unique in its stimulus and annotations.

The number of nodes constituting each meta-analytical network was different between networks used in this study. Recently, the influence of network size on single subject identifiability based on NV data has been investigated (Kröll et al., 2023), indicating that the number of nodes in a network are a relevant factor in network-level analyses. The networks that were used in this study are based on meta-analysis and represent various cognitive and psychological domains, so that network size is inherent to each network and cannot be adapted at will.

This study used the preprocessed data made available by the original authors of the dataset (Hanke et al., 2016). We acknowledge that further preprocessing steps, such as scrubbing, might influence the results. However, data quality control of the original dataset authors revealed very few motion artifacts, highlighting the beneficial effect of movie watching on participant motion (Hanke et al., 2016).

4.5 | Conclusion

The present study is the first to investigate inter- and intra-subject similarity in NFC across an FNM. Our results show that inter- and intra-subject similarity in NFC were sensitive to the progressing narrative and emotions portrayed in the movie. The ER network displayed the lowest similarity within and between subjects in NFC, followed by networks associated with long-term memory processing. The sensitivity of these networks to the FNM might be explained by the highly emotional narrative and continuous references to real world historical events, highlighting the importance of specific features and content of the chosen movie stimulus. The overarching narrative gives a unique possibility to study emotions in a social context and how they develop over time. These socio-cognitive aspects seem to specifically influence similarity within subjects, as low intra-subject similarity was additionally seen in networks involved in self- and other-related cognition. Altogether, these results show that a network perspective might help to elucidate the effects of different movie stimuli on specific cognitive domains. Additionally, the relevance of employing FNM for studying individual differences was highlighted. However, these results need to be expanded upon using different stimuli, datasets and annotations in

the future to generalize our findings. Characterizing movie stimuli in more detail to explore the effects of different features on inter- and intra-subject similarity is critical for future research in NV.

ACKNOWLEDGMENTS

The code to divide inter observer agreement time series according to the movie segments was downloaded from and is available at https://osf.io/tzpdf/. The work was supported by Deutsche Forschungsgemeinschaft (DFG), National Institute of Mental Health (R01-MH074457), Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain," European Union's Horizon 2020 Research, Innovation, Programme under Grant Agreement No. 945539 (HBP SGA3). Open access publication funded by the DFG—491111487. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The neuroimaging data were downloaded from and is openly available at https://github.com/psychoinformatics-de/studyforrest-data-phase2. The code used by the original authors of the dataset to minimally preprocess the neuroimaging data can be found under https://github.com/psychoinformatics-de/studyforrest-data-aligned/tree/master/code. The valence and arousal inter observer agreement measures were downloaded from and are available at https://f1000research.com/articles/4-92/v1#DS0.

ORCID

Lisa N. Mochalski https://orcid.org/0000-0002-9558-8251
Jean-Philippe Kröll https://orcid.org/0000-0002-2957-4468

REFERENCES

Aaron, R. V., Snodgress, M. A., Blain, S. D., & Park, S. (2018). Affect labeling and other aspects of emotional experiences in relation to alexithymia following standardized emotion inductions. *Psychiatry Research*, *262*-(July 2017), 115–123. https://doi.org/10.1016/j.psychres.2018.02.014

Aldama, F. L. (2015). The science of storytelling: Perspectives from cognitive science, neuroscience, and the humanities. *Projections*, *9*(1), 80–95. https://doi.org/10.3167/proj.2015.090106

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., … Milham, M. P. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, *4*, 170181. https://doi.org/10.1038/sdata.2017.181

Amft, M., Bzdok, D., Laird, A. R., Fox, P. T., Schilbach, L., & Eickhoff, S. B. (2015). Definition and characterization of an extended social-affective default network. *Brain Structure and Function*, *220*(2), 1031–1049. https://doi.org/10.1007/s00429-013-0698-0

Andric, M., Goldin-Meadow, S., Small, S. L., & Hasson, U. (2016). Repeated movie viewings produce similar local activity patterns but different network configurations. *NeuroImage*, *142*, 613–627. https://doi.org/10.1016/j.neuroimage.2016.07.061

Baranowski, A., & Hecht, H. (2017). One hundred years of photoplay: Hugo Münsterberg's lasting contribution to cognitive movie psychology. *Pro*, *11*(2), 1–21. https://doi.org/10.3167/proj.2017.110202

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory Jeffrey. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001.The

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., Weber, J., & Ochsner, K. N. (2014). Cognitive reappraisal of emotion: A meta-analysis of human neuroimaging studies. *Cerebral Cortex*, *24*(11), 2981–2990. https://doi.org/10.1093/cercor/bht154

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, *217*(4), 783–796. https://doi.org/10.1007/s00429-012-0380-y

Camilleri, J. A., Müller, V. I., Fox, P., Laird, A. R., Hoffstaedter, F., Kalenscher, T., & Eickhoff, S. B. (2018). Definition and characterization of an extended multiple-demand network. *NeuroImage*, *165*(October 2017), 138–147. https://doi.org/10.1016/j.neuroimage.2017.10.020

Carvalho, S., Leite, J., Galdo-Álvarez, S., & Gonçalves, Ó. F. (2012). The emotional movie database (EMDB): A self-report and psychophysiological study. *Applied Psychophysiology Biofeedback*, *37*(4), 279–294. https://doi.org/10.1007/s10484-012-9201-6

Caspers, S., Zilles, K., Laird, A. R., & Eickhoff, S. B. (2010). ALE meta-analysis of action observation and imitation in the human brain. *NeuroImage*, *50*(3), 1148–1167. https://doi.org/10.1016/j.neuroimage.2009.12.112

Chang, L. J., Jolly, E., Cheong, J. H., Rapuano, K. M., Greenstein, N., Chen, P. H. A., & Manning, J. R. (2021). Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Science. Advances*, *7*(17), eabf7129. https://doi.org/10.1126/sciadv.abf7129

Cieslik, E. C., Mueller, V. I., Eickhoff, C. R., Langner, R., & Eickhoff, S. B. (2015). Three key regions for supervisory attentional control: Evidence from neuroimaging meta-analyses. *Neuroscience and Biobehavioral Reviews*, *48*, 22–34. https://doi.org/10.1016/j.neubiorev.2014.11.003

Cutting, J. E. (2016). Narrative theory and the dynamics of popular movies. *Psychonomic Bulletin and Review*, *23*(6), 1713–1743. https://doi.org/10.3758/s13423-016-1051-4

Di Oleggio, V., Castello, M., Halchenko, Y. O., Guntupalli, J. S., Gors, J. D., & Gobbini, M. I. (2017). The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/s41598-017-12559-1

Dziura, S. L., Merchant, J. S., Alkire, D., Rashid, A., Shariq, D., Moraczewski, D., & Redcay, E. (2021). Effects of social and emotional context on neural activation and synchrony during movie viewing. *Human Brain Mapping*, *42*(18), 6053–6069. https://doi.org/10.1002/hbm.25669

Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., & Fox, P. T. (2012). Activation likelihood estimation revisited. *NeuroImage*, *59*(3), 2349–2361. https://doi.org/10.1016/j.neuroimage.2011.09.017.Activation

Eickhoff, S. B., & Grefkes, C. (2011). Approaches for the integrated analysis of structure, function and connectivity of the human brain. *Clinical EEG and Neuroscience*, *42*(2), 107–121. https://doi.org/10.1177/155005941104200211

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, *30*(9), 2907–2926. https://doi.org/10.1002/hbm.20718

Eickhoff, S. B., Milham, M., & Vanderwal, T. (2020). Towards clinical applications of movie fMRI. *NeuroImage*, *217*, 116860. https://doi.org/10.1016/j.neuroimage.2020.116860

Etkin, A., Büchel, C., & Gross, J. J. (2015). The neural bases of emotion regulation. *Nature Reviews Neuroscience*, *16*(11), 693–700. https://doi.org/10.1038/nrn4044

Finn, E. S., & Bandettini, P. A. (2021). Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *NeuroImage*, *235*(March), 117963. https://doi.org/10.1016/j.neuroimage.2021.117963

Finn, E. S., Scheinost, D., Finn, D. M., Shen, X., Papademetris, X., & Constable, R. T. (2017). Can brain state be manipulated to emphasize individual differences in functional connectivity? *NeuroImage*, *160*-(March), 140–151. https://doi.org/10.1016/j.neuroimage.2017.03.064

Fox, K. C. R., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., & Christoff, K. (2015). The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage*, *111*, 611–621. https://doi.org/10.1016/j.neuroimage.2015.02.039

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9673–9678. https://doi.org/10.1073/pnas.0504136102

Gilman, T. L., Shaheen, R., Nylocks, K. M., Halachoff, D., Chapman, J., Flynn, J. J., Matt, L. M., & Coifman, K. G. (2017). A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation. *Behavior Research Methods*, *49*(6), 2061–2082. https://doi.org/10.3758/s13428-016-0842-x

Grall, C., & Finn, E. S. (2022). Leveraging the power of media to drive cognition: A media-informed approach to naturalistic neuroscience. *Social Cognitive and Affective Neuroscience*, *17*(6), 598–608. https://doi.org/10.1093/scan/nsac019

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*(1), 1–26. https://doi.org/10.1080/1047840X.2014.940781

Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition and Emotion*, *9*(1), 87–108. https://doi.org/10.1080/02699939508408966

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., Nigbur, R., Waite, A. Q., Baumgartner, F., & Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, *3*, 160092. https://doi.org/10.1038/sdata.2016.92

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*, 140003. https://doi.org/10.1038/sdata.2014.3

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The neuroscience of film. *Pro*, *2*(1), 1–26. https://doi.org/10.3167/proj.2008.020102

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*(1), 40–48. https://doi.org/10.1016/j.tics.2009.10.011

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*-(MARCH), 1634–1640. https://doi.org/10.1126/science.1089506

Igelström, K. M., & Graziano, M. S. A. (2017). The inferior parietal lobule and temporoparietal junction: A network perspective. *Neuropsychologia*, *105*(December 2016), 70–83. https://doi.org/10.1016/j.neuropsychologia.2017.01.001

Jääskeläinen, I. P., Koskentalo, K., Balk, M. H., Autti, T., Kauramäki, J., Pomren, C., & Sams, M. (2008). Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. *The Open Neuroimaging Journal*, *2*(1), 14–19. https://doi.org/10.2174/1874440000802010014

Jenkins, L. M., & Andrewes, D. G. (2012). A new set of standardised verbal and non-verbal contemporary film stimuli for the elicitation of emotions. *Brain Impairment*, *13*(2), 212–227. https://doi.org/10.1017/BrImp.2012.18

Kauttonen, J., Hlushchuk, Y., Jääskeläinen, I. P., & Tikka, P. (2018). Brain mechanisms underlying cue-based memorizing during free viewing of movie memento. *NeuroImage*, *172*, 313–325. https://doi.org/10.1016/j.neuroimage.2018.01.068

Kreibig, S. D., & Gross, J. J. (2017). Understanding mixed emotions: Paradigms and measures. *Current Opinion in Behavioral Sciences*, *15*, 62–71. https://doi.org/10.1016/j.cobeha.2017.05.016

Kreibig, S. D., Samson, A. C., & Gross, J. J. (2013). The psychophysiology of mixed emotional states. *Psychophysiology*, *50*(8), 799–811. https://doi.org/10.1111/psyp.12064

Kreibig, S. D., Samson, A. C., & Gross, J. J. (2015). The psychophysiology of mixed emotional states: Internal and external replicability analysis of a direct replication study. *Psychophysiology*, *52*(7), 873–886. https://doi.org/10.1111/psyp.12425

Kröll, J. P., Friedrich, P., Li, X., Patil, K. R., Mochalski, L., Waite, L., Qian, X., Chee, M. W., Zhou, J. H., Eickhoff, S., & Weis, S. (2023). Naturalistic viewing increases individual identifiability based on connectivity within functional brain networks. *NeuroImage*, *273*, 120083. https://doi.org/10.1016/j.neuroimage.2023.120083

Kuppens, P., Tuerlinckx, F., Yik, M., Koval, P., Coosemans, J., Zeng, K. J., & Russell, J. A. (2017). The relation between valence and arousal in subjective experience varies with personality and culture. *Journal of Personality*, *85*(4), 530–542. https://doi.org/10.1111/jopy.12258

Labs, A., Reich, T., Schulenburg, H., Boennen, M., Mareike, G., Golz, M., Hartigs, B., Hoffmann, N., Keil, S., Perlow, M., Peukmann, A. K., Rabe, L. N., von Sobbe, F. R., & Hanke, M. (2015). Portrayed emotions in the movie "Forrest Gump". *F1000Research*, *4*, 92. https://doi.org/10.12688/f1000research.6230.1

Langner, R., & Eickhoff, S. B. (2013). Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. *Psychological Bulletin*, *139*(4), 870–900. https://doi.org/10.1037/a0030694

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M. Y., Gilmore, A. W., McDermott, K. B., Nelson, S. M., Dosenbach, N. U., Schlaggar, B. L., Mumford, J. A., Poldrack, R. A., & Petersen, S. E. (2015). Functional system and areal organization of a highly sampled individual human brain. *Neuron*, *87*(3), 657–670. https://doi.org/10.1016/j.neuron.2015.06.037

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Lettieri, G., Handjaras, G., Ricciardi, E., Leo, A., Papale, P., Betta, M., Pietrini, P., & Cecchetti, L. (2019). Emotionotopy in the human right temporo-parietal cortex. *Nature Communications*, *10*(1), 5568. https://doi.org/10.1038/s41467-019-13599-z

Lettieri, G., Handjaras, G., Setti, F., Cappello, E. M., Bruno, V., Diano, M., Leo, A., Ricciardi, E., Pietrini, P., & Cecchetti, L. (2022). Default and control network connectivity dynamics track the stream of affect at multiple timescales. *Social Cognitive and Affective Neuroscience*, *17*(5), 461–469. https://doi.org/10.1093/scan/nsab112

Liu, X., Hairston, J., Schrier, M., & Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages. *Neuroscience & Biobehavioral Reviews*, *35*(5), 1219–1236. https://doi.org/10.1016/j.neubiorev.2010.12.012.Common

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., Jefferies, E., & Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(44), 12574–12579. https://doi.org/10.1073/pnas.1608282113

Morawetz, C., Bode, S., Baudewig, J., & Heekeren, H. R. (2017). Effective amygdala-prefrontal connectivity predicts individual differences in successful emotion regulation. *Social Cognitive and Affective Neuroscience*, *12*(4), 569–585. https://doi.org/10.1093/scan/nsw169

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social*

Cognitive and Affective Neuroscience, *14*(6), 669–687. https://doi.org/10.1093/scan/nsz037

Nguyen, V. T., Sonkusare, S., Stadler, J., Hu, X., Breakspear, M., & Guo, C. C. (2017). Distinct cerebellar contributions to cognitive-perceptual dynamics during natural viewing. *Cerebral Cortex*, *27*(12), 5652–5662. https://doi.org/10.1093/cercor/bhw334

Noble, S., Scheinost, D., Finn, E. S., Shen, X., Papademetris, X., McEwen, S. C., Bearden, C. E., Addington, J., Goodyear, B., Cadenhead, K. S., Mirzakhanian, H., Cornblatt, B. A., Olvet, D. M., Mathalon, D. H., McGlashan, T. H., Perkins, D. O., Belger, A., Seidman, L. J., Thermenos, H., … Constable, R. T. (2017). Multisite reliability of MR-based functional connectivity. *NeuroImage*, *146*, 959–970. https://doi.org/10.1016/j.neuroimage.2016.10.020

Nostro, A. D., Müller, V. I., Varikuti, D. P., Pläschke, R. N., Hoffstaedter, F., Langner, R., Patil, K. R., & Eickhoff, S. B. (2018). Predicting personality from network-based resting-state functional connectivity. *Brain Structure & Function*, *223*(6), 2699–2719. https://doi.org/10.1007/s00429-018-1651-z

Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., & Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(24), 9599–9604. https://doi.org/10.1073/pnas.1206095109

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage*, *211*(November 2019), 116604. https://doi.org/10.1016/j.neuroimage.2020.116604

Pläschke, R. N., Cieslik, E. C., Müller, V. I., Hoffstaedter, F., Plachti, A., Varikuti, D. P., Goosses, M., Latz, A., Caspers, S., Jockwitz, C., Moebus, S., Gruber, O., Eickhoff, C. R., Reetz, K., Heller, J., Südmeyer, M., Mathys, C., Caspers, J., Grefkes, C., … Eickhoff, S. B. (2017). On the integrity of functional brain networks in schizophrenia, Parkinson's disease, and advanced age: Evidence from connectivity-based single-subject classification. *Human Brain Mapping*, *38*(12), 5845–5858. https://doi.org/10.1002/hbm.23763

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2011). Functional network organization of the human brain. *Neuron*, *72*(4), 665–678. https://doi.org/10.1016/j.neuron.2011.09.006

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *Neuroimage*, *216*, 116392. https://doi.org/10.12688/f1000research.6230.1

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., Fox, P. T., & Eickhoff, S. B. (2012). Modelling neural correlates of working memory: A coordinate-based meta-analysis. *NeuroImage*, *60*(1), 830–846. https://doi.org/10.1016/j.neuroimage.2011.11.050

Saarimäki, H. (2021). Naturalistic stimuli in affective neuroimaging: A review. *Frontiers in Human Neuroscience*, *15*, 675068. https://doi.org/10.3389/fnhum.2021.675068

Sabatinelli, D., Fortune, E. E., Li, Q., Siddiqui, A., Krafft, C., Oliver, W. T., Beck, S., & Jeffries, J. (2011). Emotional perception: Meta-analyses of face and natural scene processing. *NeuroImage*, *54*(3), 2524–2533. https://doi.org/10.1016/j.neuroimage.2010.10.011

Samson, A. C., Kreibig, S. D., Soderstrom, B., Wade, A. A., & Gross, J. J. (2016). Eliciting positive, negative and mixed emotional states: A film library for affective scientists. *Cognition and Emotion*, *30*(5), 827–856. https://doi.org/10.1080/02699931.2015.1031089

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., Yeo, B. T. T., & Yeo, B. T. T. (2017). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, *28*, 1–20. https://doi.org/10.1093/cercor/bhx179

Schaefer, A., Nils, F., Philippot, P., & Sanchez, X. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, *24*(7), 1153–1172. https://doi.org/10.1080/02699930903274322

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Sengupta, A., Kaule, F. R., Guntupalli, J. S., Hoffmann, M. B., Häusler, C., Stadler, J., & Hanke, M. (2016). A studyforrest extension, retinotopic mapping and localization of higher visual areas. *Scientific Data*, *3*, 1–14. https://doi.org/10.1038/sdata.2016.93

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., Beckmann, C. F., & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, *106*, 13040–13045. https://doi.org/10.1073/pnas.0905267106

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences*, *23*, 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Spreng, R. N., & Grady, C. L. (2010). Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. *Journal of Cognitive Neuroscience*, *22*(6), 1112–1123. https://doi.org/10.1162/jocn.2009.21282

Tarvainen, J., Westman, S., & Oittinen, P. (2015). The way films feel: Aesthetic features and mood in film. *Psychology of Aesthetics, Creativity, and the Arts*, *9*(3), 254–265. https://doi.org/10.1037/a0039432

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*(October 2018), 100600. https://doi.org/10.1016/j.dcn.2018.10.004

Vanderwal, T., Eilbott, J., Finn, E. S., Craddock, R. C., Turnbull, A., & Castellanos, F. X. (2017). Individual differences in functional connectivity during naturalistic viewing conditions. *NeuroImage*, *157*(June), 521–530. https://doi.org/10.1016/j.neuroimage.2017.06.027

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., & Castellanos, F. X. (2015). Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage*, *122*, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069

Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta- analysis RAINER. *European Journal of Social Psychology*, *26*, 557–580.

Witt, S. T., Meyerand, M. E., & Laird, A. R. (2008). Functional neuroimaging correlates of finger tapping task variations: An ALE meta-analysis. *NeuroImage*, *71*(2), 233–236. https://doi.org/10.1038/mp.2011.182.doi

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Yik, M. S. M., Russell, J. A., & Barrett, L. F. (1999). Structure of self-reported current affect: Integration and beyond. *Journal of Personality and Social Psychology*, *77*(3), 600–619. https://doi.org/10.1037/0022-3514.77.3.600

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mochalski, L. N., Friedrich, P., Li, X., Kröll, J.-P., Eickhoff, S. B., & Weis, S. (2024). Inter- and intra-subject similarity in network functional connectivity across a full narrative movie. *Human Brain Mapping*, *45*(11), e26802. https://doi.org/10.1002/hbm.26802

5 Test–Retest Reliability of Meta Analytic Networks During Naturalistic Viewing, Jean-Philippe Kröll, Patrick Friedrich, Xuan Li, Yulia Nurislamova, Nevena Kraljevic, Anna Geiger, Julia Mans, Laura Waite, Julian Caspers, Xing Qian, Michael WL Chee, Juan Helen Zhou, Simon Eickhoff, Susanne Weis, ***bioRxiv, 2024.05.15.594266 doi: https://doi.org/10.1101/2024.05.15.594266***

Test–Retest Reliability of Meta Analytic Networks During Naturalistic Viewing

Jean-Philippe Kröll [a, b, +], Patrick Friedrich [a, b], Xuan Li [a, b], Yulia Nurislamova [a, b], Nevena Kraljevic[a, b], Anna Geiger [a, b], Julia Mans [a, b], Laura Waite [a], Julian Caspers [c],Xing Qian[e], Michael WL Chee [d,f], Juan Helen Zhou [d,e,f], Simon Eickhoff [a, b], Susanne Weis [a, b]

a *Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich 52428, Germany*

b *Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany*

c Department of Diagnostic and Interventional Radiology, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

d Centre for Sleep and Cognition & Centre for Translational MR Research, Yong Loo Lin School of Medicine, National University of Singapore

e Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

f Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore, Singapore

[+] *corresponding author , email: j.kroell@fz-juelich.de*

Abstract

Functional connectivity analyses have given considerable insights into human brain function and organization. As research moves towards clinical application, test-retest reliability has become a main focus of the field. So far, the majority of studies have relied on resting-state paradigms to examine brain connectivity, based on its low demand and ease of implementation. However, the reliability of resting-state measures is mostly moderate, potentially due to its unconstrained nature. Recently, naturalistic viewing paradigms have gained popularity because they probe the human brain under more ecologically valid conditions, thereby possibly increasing reliability. Therefore, we here compared the reliability of graph metrics extracted from resting-state and naturalistic viewing in functional networks, across two sessions. We show that naturalistic viewing can increase reliability over resting-state, but that its effect varies between stimuli and networks. Furthermore, we demonstrate that the effect of naturalistic viewing differs between two cohorts with Asian and European cultural backgrounds. Taken together, our study encourages the use of naturalistic viewing to increase reliability, but emphasizes the need to carefully select the appropriate stimulus and network for the respective research question.

Introduction

Functional magnetic resonance imaging (fMRI) data has become a widely-used tool to investigate neurological diseases and their underlying patterns (Balthazar et al., 2014; Basaia et al., 2019; Supekar et al., 2008; Wu et al., 2009). The critical assumption behind all of these studies is that the measured brain activity is reliable, such that differences between subjects and timepoints are interpretable. However, the reported reliability of fMRI measures varies vastly across studies (Bennett and Miller, 2010), due to small test-retest samples and different analysis choices. As fMRI research moves towards the identification of biomarkers (Bassett et al., 2008; Rubinov et al., 2009; Supekar et al., 2008; Wang et al., 2009), increasing reliability has become a priority. In order to aid in the diagnosis and prognosis of brain disorders, a measure has to be capable of giving consistent results, otherwise it is unsuitable as a biomarker.

The majority of prior reliability studies have relied on metrics derived from resting state (RS) (Braun et al., 2012; Guo et al., 2012; Wang et al., 2011). With low demands on the participants, RS is well suited for healthy as well as patient cohorts and allows for a quick data acquisition. Although various studies reported moderate to good reliability of RS-derived measures (Braun et al., 2012; Deuker et al., 2009; Wang et al., 2011), the RS paradigm also suffers from a few drawbacks. Data acquired during the RS can be strongly confounded by head movement and drowsiness of the participant due to its unconstrained nature (Tagliazucchi and Laufs, 2014; Van Dijk et al., 2012), as participants struggle to remain awake and motionless in the absence of a task or stimulus. For the same reasons, RS is more susceptible to be influenced by spontaneous thought of the participant (Christoff et al., 2004; Gonzalez-Castillo et al., 2021).
Naturalistic Viewing (NV) paradigms, during which participants are presented with a story or a film, have recently gained popularity because they might give insight into the brain's function under more ecologically valid conditions. It has been shown that NV poses several advantages over conventional RS such as increased participant engagement, reduced head movement and increased synchronization between subjects (Hasson et al., 2004; Wang et al., 2017). Especially relevant for clinical studies, NV shares with RS the advantage of minimizing demand on the participants (Eickhoff et al., 2020). On the other hand, NV paradigms place a behavioral constraint that allows for the study of normal and abnormal brain function, somewhat similar to task-based designs. Making use of these advantages, a series of studies could show altered connectivity during NV in patients (Guo et al., 2016, 2015; Hyett et al., 2015; Yang et al., 2020), encouraging the application of NV measures as biomarkers.

Furthermore, several studies suggest that NV increases test-retest reliability in comparison with RS (O'Connor et al., 2017; Wang et al., 2017; Zhang et al., 2022). This improvement can be attributed to several factors. First, many studies have pointed out that NV improves signal properties by increasing participant engagement (Eickhoff et al., 2020; Finn and Bandettini, 2020; Li et al., 2022; Vanderwal et al., 2017). Secondly, by reducing head movement and drowsiness, NV is less susceptible to noise than conventional RS. Thirdly, by presenting the same stimulus across sessions, NV is less influenced by spontaneous thought of the participant while also placing a behavioral constraint that reduces variance. However, the effect of NV on reliability is dependent on various factors such as attention (Ki et al., 2016), successful episodic encoding (Hasson et al., 2008) as well as the chosen movie stimulus (Hasson et al., 2010; Kröll et al., 2023; Tian et al., 2021) and differs between different brain regions and networks (Wang et al., 2017).

The present study aims to further evaluate the test-retest reliability of NV, by investigating its influence on the reliability of five commonly used graph theoretical measures. The application of graph theoretical measures to fMRI data is an established method (Braun et al., 2012; Guo et al., 2012; Reijneveld et al., 2007; Stam and Reijneveld, 2007), and has given insights into the complex functional structure of the brain (Bullmore and Sporns, 2009; Rubinov and Sporns, 2010), both in healthy and patient cohorts. To benchmark the reliability of NV, we compare it to that of RS. Further, we evaluate the influence of the movie content, by employing stimuli with different levels of social content, ranging from the neutral movie *Inscapes*, over the silent movie *The Circus*, to the most social movie *Indiana Jones and the Temple of Doom*. In addition, several authors have suggested that the same NV stimuli might deviate in its effect between different populations (Eickhoff et al., 2020; Hasson et al., 2010; Telesford et al., 2010). The cultural background of a participant is likely to influence how a given movie is perceived and might result in deviating effects across cohorts. Therefore, in this study, we compare the effect of NV in two independent samples from Europe and Asia, respectively, using the same stimuli. In contrast to the majority of previous studies, we here compare reliability on the basis of a priori defined networks, and not on a whole-brain basis.
The analysis of network based measures allows us to investigate how NV influences the reliability in different cognitive domains. The networks implemented in this study are meta-analytically defined networks that represent the most likely core nodes involved in a given cognitive function, because they incorporate convergent information from a multitude of studies.

2. Methods

2.1 Participants

Dataset IMAX
For the first dataset, 36 healthy right-handed and ambidextrous adults were scanned at the Centre for Translational MR Research, National University of Singapore. Exclusion criteria were neurological or psychiatric diagnoses, significant visual or hearing impairment, alcohol or caffeine consumption 6 hours prior to the scan and self-reporting of bad sleep the night before the scan days. All participants underwent three identical testing sessions within a one-week interval. Subjects gave written, informed consent and were compensated for their participation. The study was approved by the institutional review board of the National University of Singapore.

Dataset JUMAX
For the second dataset, 36 healthy adults were scanned at the Forschungszentrum Jülich. Exclusion criteria were neurological or psychiatric diagnoses, significant visual or hearing impairment, alcohol or caffeine consumption 6 hours prior to the scan and self-reporting of bad sleep the night before the scan days. All participants underwent three identical testing sessions within a one-week interval. Subjects gave written, informed consent and were compensated for their participation. The study was approved by the ethics committee of the Heinrich Heine University, Düsseldorf.

Due to unavailability of part of the data of the JUMAX sample, the final cohort comprised 33 subjects (14 females, mean age 27.5 +/- 3 years). Accordingly, to match the number of available subjects from the JUMAX dataset, only the first 33 subjects were used from the IMAX sample (17 females, 27 +/- 2.7). For all subsequent analyses, only the first two sessions of both samples were used.

2.2 Data acquisition

For both datasets, the data was acquired on a Siemens Magnetom PrismaFit 3-Tesla with a 20-Channel head coil. Structural images were collected using an MP-RAGE sequence (TR=2300ms, TE =2,28ms, TI=900ms, flip-angle=8°) and 1mm voxel size. All RS and NV runs used the same echo planar imaging sequence (TR=719ms, TE=30ms, flip-angle=52°, slices=44, FOV=225x225 mm2) resulting in 2.96x2.96x3 mm voxel size. Data from collaborators at the National University of Singapore were retrieved and structured in the form of a DataLad dataset, a research data management solution providing data versioning, data transport, and provenance capture (Halchenko et al., 2021). Each of the three testing sessions per participant, which were conducted within a seven day period, comprised three NV runs and two RS scans. The order of scans was identical on all three days, starting with a structural scan, followed by 5 functional scans in the order of RS 1, Inscapes, Circus, Indiana Jones and RS 2, with each functional scan lasting for 10 minutes. All movies had been cut to the same length. For RS scans, participants were asked to lay as still as possible and think of nothing in particular, while keeping their eyes open. Instructions for the NV scans were to watch the movies while staying as still as possible. For all scans, participants were asked to not fall asleep during the measurement. Foam wedges were fitted around each subject's head for comfort and to decrease movement. For all subsequent analyses, only the first two scan sessions and the first RS scan (RS1) of each session were used. The movie clips were presented via a mirror that was mounted on the head coil and the sound was played through headphones.

2.3 Stimulus material

Three different movie stimuli with different levels of social content (Inscapes < The Circus < Indiana Jones) were used. Inscapes is a nonverbal, non-social series of animated abstract shapes created by Vanderwal et al. which was looped to match the 10 minutes duration (Vanderwal et al., 2015). The Circus (United Artists Digital Studios, 1928, directed by Charlie Chaplin) is a silent black-and-white. Participants were shown the first 10 minutes of the film which depicts the protagonist being chased by the police and unintentionally causing comic situations during his escape. Indiana Jones and the Temple of Doom (Paramount Pictures, 1984, directed by Steven Spielberg) shows the first 10 minutes of the movie during which the protagonist has to fight off several hitmen who are trying to kill him and finally escapes by taking a plane. The end of the clips used from The Circus and Indiana Jones both coincide with a change of scene in the respective movie itself.

2.4 Data preprocessing

Preprocessing of MRI data was performed using fMRIPrep, version 22.0.0 (Esteban et al., 2019). In brief, the T1-weighted volumes were corrected for intensity non-uniformity and skull-stripped. The extracted brain images were then transformed into Montreal Neurological Institute (MNI) space and motion corrected using Advanced Normalization Tools (Avants et

al., 2009). The functional data was motion-corrected with MCflirt (Jenkinson et al., 2002) and subsequently co-registered to the native T1-weighted image using boundary based registration with six degrees of freedom from Freesurfer (Greve and Fischl, 2009). Subsequently, an isotropic Gaussian kernel of 6mm FWHM (full-width half-maximum) was applied for spatial smoothing. The images were further regressed out of nuisance signals and bandpass filtered (0.01– 0.1 Hz). Nuisance signals were the global signals extracted within the CSF, the WM, and the whole-brain masks which were regressed from the preprocessed fMRI data for each subject. In addition, the standard six motion parameters and their first temporal derivatives were regressed out.
Subsequently, network functional connectivity (NFC) matrices were constructed for 14 meta-analytical networks, comprising nine to 23 nodes (a detailed description of the networks can be found in the supplements). In short, isotropic 5 mm spheres were created around the local maxima of each meta-analytical network node and only gray matter voxels were included. Using the Junifer toolbox (Synchon Mandal et al., 2023), we extracted the mean time series of each node and computed the Pearson's correlation coefficient between all node pairs to produce a node times node connectivity matrix for each subject and each condition. The networks cover affective (Amft et al., 2015; Buhle et al., 2014; Liu et al., 2011; Sabatinelli et al., 2011), social (Amft et al., 2015; Bzdok et al., 2012; Caspers et al., 2010), executive(Camilleri et al., 2018; Cieslik et al., 2015; Langner and Eickhoff, 2013; Rottschy et al., 2012), memory (Binder et al., 2009; Spreng et al., 2009) and motor (Witt et al., 2008) functions.

2.5 Graph theoretical analyses

Subsequently, graph metrics were derived from the NFC matrices. The fully connected node x node matrices were thresholded at 0.1 to determine the presence or absence of connections (edges) between nodes. Connections above the threshold retained their correlation coefficient, whereas subthreshold edges were assigned values of 0. This thresholding procedure was performed on both positive and negative connections.
Five different Graph metrics were extracted from the thresholded NFC matrices using the *NetworkX* toolbox (A Hagberg et al., 2008), including degree centrality, clustering coefficient, betweenness centrality, global efficiency and mean shortest path length. Degree centrality measures the connectedness of each node, computed as the weighted sum of all edges connected to that node. The clustering coefficient for a given node is a measure of local connectedness, measuring the proportion of existing connections out of all possible connections between the nearest neighbors of that node. Betweenness centrality measures the centrality of a node in the network, calculated as the ratio of shortest paths (that is the smallest number of links that need to be traversed to go from one node to another) in the whole graph that pass through that node. The efficiency of a pair of nodes in a graph is the reciprocal of the shortest path distance between these two nodes. The global efficiency of a graph is the average efficiency of all pairs of nodes. Shortest path length denotes the minimum number of nodes that need to be passed through to connect one node to another. Mean shortest path length is the average shortest path length between all nodes of the graph.

2.6 Test-retest reliability

The reliability of each graph metric was quantified by calculating the intraclass correlation coefficient (ICC) across these measures derived from the two scans (McGraw and Wong, 1996; Shrout and Fleiss, 1979). A one-way ANOVA was applied to the measures of the two scan sessions across subjects, to calculate between-subject mean square (MSp) and mean square error (MSe). ICC values were then calculated as:

$$ICC = MSp - MSe \,/\, MSp + (d - 1)\, MSe$$

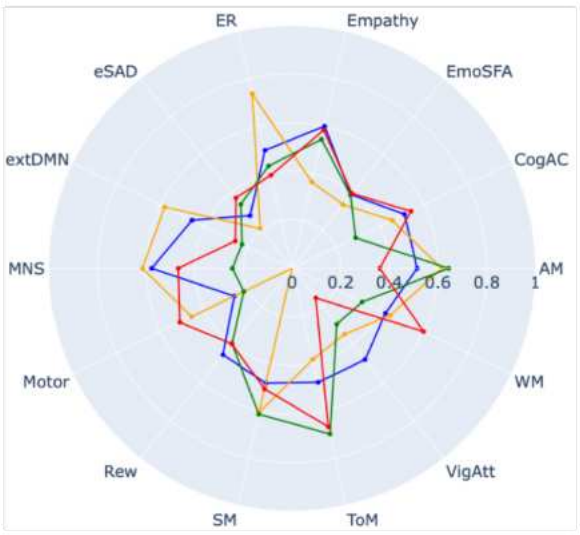
where d is equal to the number of observations per subject. For each graph measure, we calculated reliability at the scan-wise level. Scan-wise reliability estimates the reliability of one score derived from the entire scan session, opposed to calculating one ICC value for the graph metric of each node (Guo et al., 2012; Wang et al., 2017). Here, a single ICC value was calculated for the mean graph metric averaged across all nodes of the network. The reliability results are considered excellent (ICC > 0.8), good (ICC 0.6-0.79), moderate (ICC 0.4-0.59), fair (ICC 0.2-0.39), and poor (ICC < 0.2) (Guo et al., 2012). As negative ICCs are difficult to interpret and reasons for negative values are unclear (Müller and Büttner, 1994), in the following we set negative ICCs to zero (that is completely non-reliable) as has been suggested in previous studies (Braun et al., 2012; Kong et al., 2007; Zhang et al., 2011).

3. Results

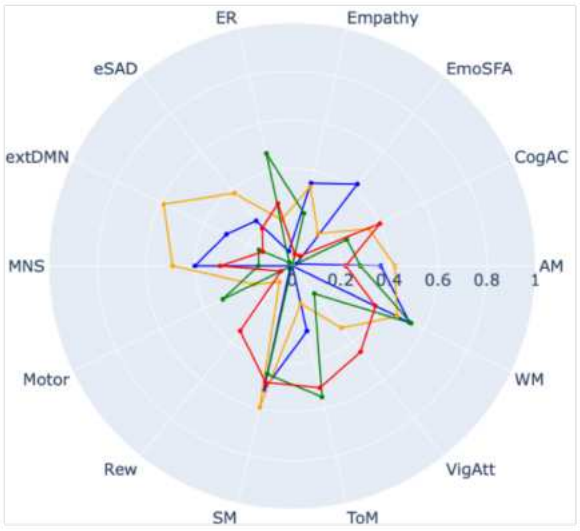
3.1 Reliability of graph metrics in the IMAX sample

Degree centrality

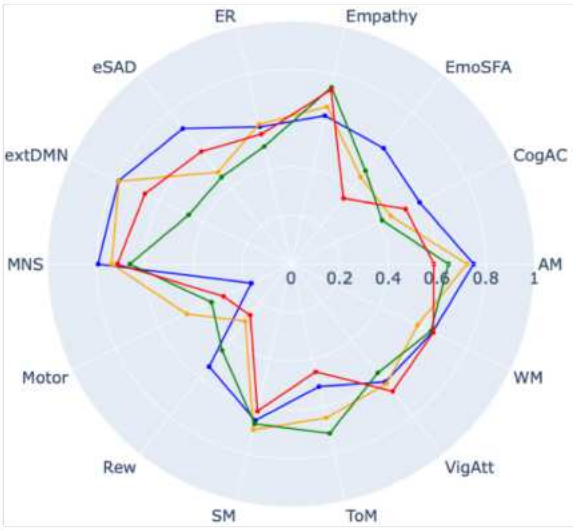
Cluster coefficient



Between centrality



Efficiency



Shortest path length

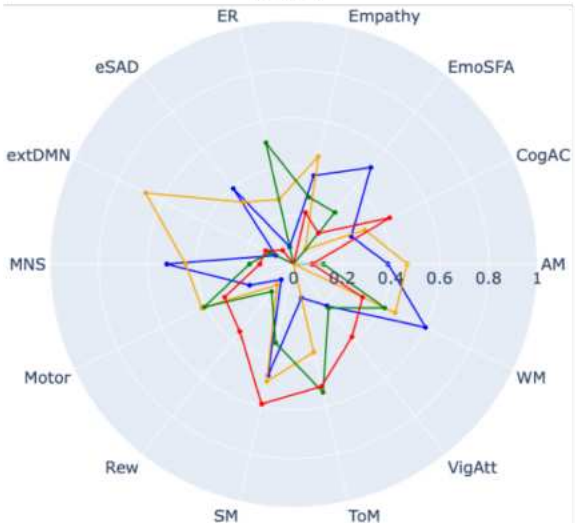


Figure 1. ICC of graph metrics across the 14 networks in the IMAX sample. Graph metrics are shown for the RS scan and three different movies. ICC values below zero are not depicted. (AM =Autobiographical Memory, CogAC = Cognitive Attention Control,eMDN=extended Multiple Demand Network, EmoSF= Emotional Scene and Face Processing, ER = Emotion Regulation, eSAD=Extended Social-affective Default, MNS = Mirror Neuron System, Rew = Reward, SM = Semantic Memory, ToM = Theory of Mind, VigAtt= Vigilant Attention, WM = Working memory)

We investigated the reliability of five graph measures derived from 14 different networks. For the IMAX sample, we found low to good reliability across networks. Degree centrality, cluster coefficient and efficiency showed a trend towards higher reliability than between centrality and shortest path length.

Degree centrality showed the highest ICC during RS in five (AM, CogAC, MNS, Rew and VigAtt), during Inscapes in three (SM, ER, extDMN), during Circus in four (EmoSF, Empathy, ToM, WM) and during Jones in three (eSAD, Motor, Empathy) networks.

Cluster coefficient showed the highest ICC during RS in four (Rew, Empathy, VigAtt, EmoSF), during Inscapes in three (MNS,ER,extDMN), during Circus in three (AM, SM, ToM) and during Jones in five (CogAC, Motor, EmoSF, eSAD, WM) networks.

Efficiency showed the highest ICC during RS in eight (AM, MNS, CogAC, EmoSF, Rew, eSAD, extDMN, WM), during Inscapes in three (Motor, SM, ER, extDMN), during Circus in three (Empathy, ToM, WM) and during Jones in two (VigAtt, WM) networks.

Between centrality showed the highest ICC during RS in two (EmoSF, Empathy), during Inscapes in five networks (AM, MNS, SM, eSAD, extDMN), during Circus in four (Motor, ER, ToM, WM) and during Jones in three (CogAC, Rew, VigAtt) networks.

Shortest path length showed the highest ICC during RS in four (MNS, EmoSF, eSAD, WM), during Inscapes in four (AM, Motor, Empathy, extDMN), during Circus in two (ER, ToM) and during Jones in four (CogAC, Rew, SM, VigAtt) networks.

3.2 Reliability of graph metrics in the JUMAX sample

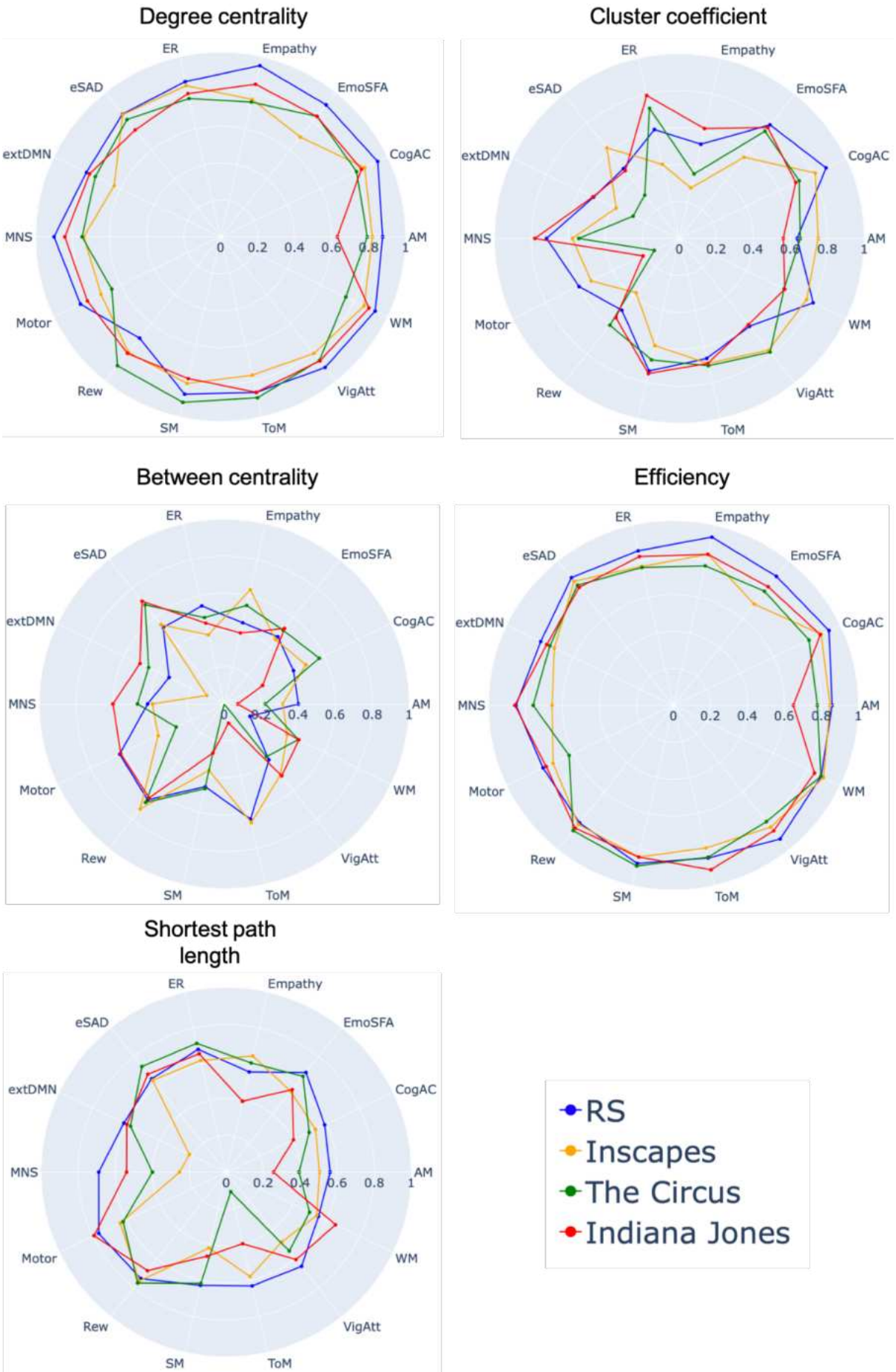


Figure 2. ICC of graph metrics across the 14 networks in the JUMAX sample. Graph metrics are shown for the RS scan and three different movies. ICC values below zero are not depicted. (AM =Autobiographical Memory, CogAC = Cognitive Attention Control,eMDN=extended Multiple Demand Network, EmoSF= Emotional Scene and Face Processing, ER = Emotion Regulation, eSAD=Extended Social-affective Default, MNS = Mirror Neuron System, Rew = Reward, SM = Semantic Memory, ToM = Theory of Mind, VigAtt= Vigilant Attention, WM = Working memory)

For JUMAX we found low to excellent reliability across networks. Degree centrality, cluster coefficient and efficiency showed a trend towards higher reliability than between centrality and shortest path length.

Degree centrality showed the highest ICC during RS in nine (AM, CogAC, EmoSF, Empathy, ER, MNS, Motor, VigAtt, WM), during Inscapes in one (eSAD) and during Circus in three (Rew, SM, ToM) networks.

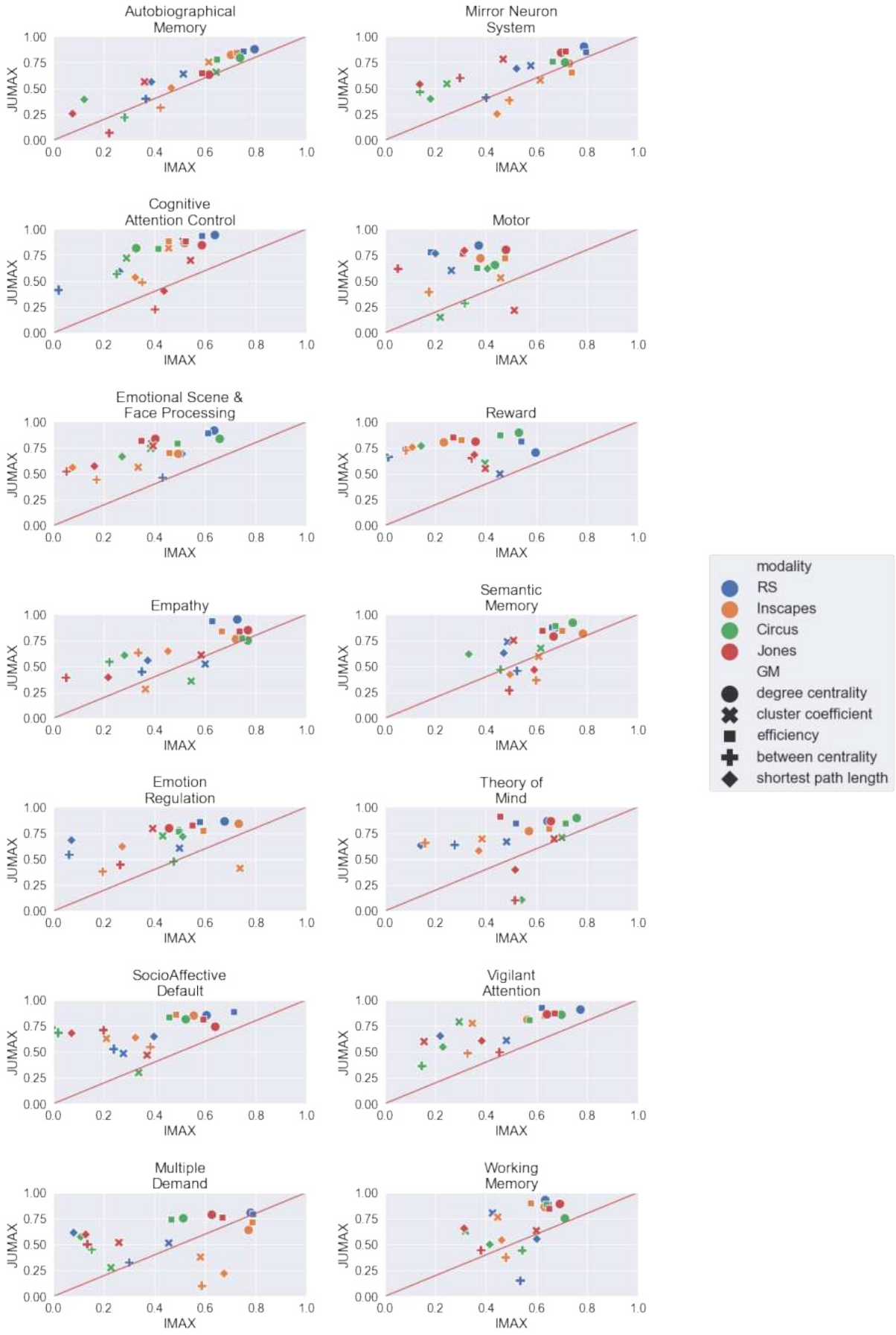
Cluster coefficient showed the highest ICC during RS in five (CogAC, Motor, EmoSF, SM, WM), during Inscapes in two (AM, eSAD), during Circus in four (Rew, ER, ToM, VigAtt) and during Jones in three (MNS, Empathy, extDMN) networks.

Efficiency showed the highest ICC during RS in ten (AM, MNS, CogAC, Motor, EmoSF, Empathy, ER, eSAD, VigAtt, extDMN), during Circus in three (Rew, SM, WM) and during Jones in one (ToM) networks.

Between centrality showed the highest ICC during RS in four (AM, Motor, Rew, ER), during Inscapes in two (Empathy, ToM), during Circus in one (SM) and during Jones in seven (CogAC, EmoSF, MNS, eSAD, VigAtt, extDMN, WM) networks.

Shortest path length showed the highest ICC during RS in nine (AM, MNS, CogAC, EmoSF, Rew, ER, SM, ToM, VigAtt, extDMN), during Circus in two (Empathy, SM) and during Jones in three (Motor, eSAD, WM) networks.

3.3 Comparison of the two samples



Comparing the results across the two samples, it was evident that the ICC was generally higher in the JUMAX sample than in the IMAX sample. However, in both samples, degree centrality and efficiency tended to show the highest ICCs, followed by cluster coefficient and then by between centrality and shortest path length. The AM, MNS, Empathy and SM networks showed similar results in both samples, while the rest of the networks showed more distinct results. Overall there was not one stimulus which led to more consistent results than other stimuli across the two samples.

4 Discussion

The primary goal of this study was to investigate the reliability of NV and RS, across various functional networks. Graph metrics indicate that NV is - in certain conditions - more reliable than RS, consistent with previous results from Wang et al. 2017 (Wang et al., 2017). However, our results demonstrated that this effect is dependent on a variety of factors. Firstly, the choice of the NV stimulus impacts the reliability of a given graph metric. Secondly, the effect of NV stimuli varies across cohorts. Thirdly, the increase in reliability is not uniform across the brain, but varies between different functional networks.

NV vs RS

Starting from observations indicating that graph metrics extracted from RS fMRI can be used to investigate abnormalities in brain organization (Petrella, 2011; Wu et al., 2009), researchers have focused on investigating the reliability with which these graph metrics can be extracted. With ongoing efforts to use characteristic abnormalities to successfully detect and track neurological diseases, it will be crucial to increase reliability as much as possible. Therefore, researchers have shifted to extracting graph metrics from other modalities than RS such as task-based fMRI (Aron et al., 2006; Cao et al., 2014) or NV (Rikandi et al., 2022; Zhang and Liu, 2021). In contrast to task-free RS, these modalities place a constraint on the participant which might reduce variability that is otherwise induced by spontaneous thoughts (Finn et al., 2017; Hasson et al., 2010; Vanderwal et al., 2017). Our results confirmed the notion that behavioral constraints can prove to be beneficial to increase reliability over unconstrained RS. In multiple networks, NV stimuli increased reliability of one or more graph metrics in comparison with RS (Fig.1, Fig.2). Furthermore, this improvement of reliability is observable across networks dealing with affective, social, executive, memory and motor functions, indicating that NV increases engagement not only in sensory, but also in higher order networks. On the other hand, our results also showed that in many instances RS was more reliable than NV, which is in line with previous studies that showed that NV does not unconditionally increase reliability (Hlinka et al., 2022; Zhang et al., 2022). Nevertheless, these results, in our opinion, encourage the use of NV to improve reliability as NV increased reliability over RS drastically in certain cases. But rather than viewing NV as a one-fits-all tool, our findings further underline the importance of using specific NV stimuli (and brain networks) for a specific purpose.

The observed reliability in our study matches results from previous studies investigating graph metrics extracted from RS and NV (Braun et al., 2012; Cao et al., 2014; Wang et al., 2017). However, in contrast to Wang et al (Wang et al., 2017) we showed that NV does not generally improve reliability of graph metrics, but that its effect varies across networks, stimuli and graph metric.

Variance across cohorts

One of the advantages of using NV stimuli is that they are easier to share across multiple sites than traditional tasks (DuPre et al., 2020; Eickhoff et al., 2020). By combining data from a multitude of studies using the same NV stimuli, one can not only achieve large sample sizes, but also place the same behavioral constraint on all subjects across sites. However, several studies have suggested that cultural differences between movies (and/or cohorts) might hinder generalizability (DuPre et al., 2020; Eickhoff et al., 2020; Hasson et al., 2010). In this study, we compared an asian and a european cohort that were subjected to the same three NV stimuli. In both samples, NV stimuli increased reliability of graph metrics in comparison with RS. However, we did not observe that the same combination of stimulus, network and graph metric led to improved reliability over RS in both samples (Fig.3). Although some of the networks (AM, MNS, Empathy and SM) show similar trends, it is not generally the case that results from both samples are highly overlapping. These differences might have been driven by the different cultural backgrounds of the participants. The appreciation of a film is culturally specific (Saarimäki, 2021) and likely different between the european and asian cohorts. Several studies have demonstrated cultural differences in the perception of faces (Adams et al., 2010; Goh et al., 2010; Harada et al., 2020), a factor that is especially relevant for the NV stimuli Circus and Jones during which a variety of different faces are depicted. Related, in a study from Sneddon et al., 2011 (Sneddon et al., 2011) participants from Northern Ireland, Serbia, Guatemala and Peru showed systematic differences in their rating of positive and negative emotions being displayed in twelve short movie clips. Our study provides further evidence for the notion that future studies should take into account cultural differences between cohorts when selecting a movie stimulus.

Variance across networks and stimuli

In our analysis, we employed meta-analytically defined networks that represent the most likely core nodes of a given brain function. Alternatively to approaches where the effect of NV is considered from a whole brain perspective, we here investigated how NV engages different networks. Similar to previous studies, we observed that the effect of different NV stimuli varies across different networks (Finn and Bandettini, 2020; Kröll et al., 2023; Wang et al., 2017) and reliability of graph metrics was not unconditionally increased over RS. One of the advantages of NV is the possibility to more effectively engage brain networks of interest, in comparison with RS (Eickhoff et al., 2020; Guo et al., 2015). Intuitively, one would expect that a network responsible for the processing of emotions is differently engaged by an emotional clip than e.g. the motor network. This effect can also be seen in our results as different networks exhibit varying reliabilities in response to the same stimulus

To analyze the effect of the chosen movie stimulus on the reliability of a given graph metric, we employed three movies with different levels of social content. Various studies have shown that different NV stimuli can lead to significantly different results. Finn et al., 2021 reported that FC derived from different movies varied in its ability to accurately predict emotion and cognition scores (Finn and Bandettini, 2020). Similarly, Gal et al., 2022 showed that the accuracy with which task-activation maps could be predicted differed between FC derived from Hollywood NV stimuli and independent NV stimuli (Gal et al., 2022). Our results extend these findings by showing that NV stimuli also divert in their impact on the reliability of extracted graph measures. Previous studies have shown that reliability is strongly dependent on attention (Ki et al., 2016) and several studies have suggested that NV stimuli with social content are best suited to engage participants and keep their attention over a longer period of time (Finn and Bandettini, 2020; Saarimäki, 2021; Schaefer et al., 2010). In

line with that, we observed a tendency that the two more social stimuli, Circus and Jones, more frequently led to improvement than the abstract movie Inscapes.
However, in the majority of cases, reliability was higher for graph metrics extracted from RS than these extracted from NV. This was somewhat unexpected since RS is generally seen as an unconstrained state and one would therefore expect more variability between sessions than for more constrained states like NV. Several factors might have led to the relative decrease in reliability for NV. Firstly, familiarity with a given movie might have played a role as multiple studies have shown that expected stimuli reduce the neuronal response (Alink et al., 2010; Koster-Hale and Saxe, 2013). The sessions for both datasets were conducted within a week and therefore participants will be familiar with the movie during the second session. This effect might have induced variability for the NV conditions, while RS on the other hand has been shown to remain stable across sessions (Mason et al., 2007; Wang et al., 2011). Secondly, some of the networks employed here (AM, SM and eSAD) are overlapping with the default mode network which is linked to intrinsically oriented functions, rather than the processing of external stimuli (Golland et al., 2007; Hasson et al., 2004). This may plausibly lead to decreased reliability of NV in comparison with RS, in these networks. These results emphasize that future studies should carefully consider which combination of graph metric, stimulus and network is suited for the research question at hand. Using purpose-built movies, such as emotionally salient clips for patients with depression (Guo et al., 2016), in combination with the functionally involved network will help improve reliability and advance the characterization of disease specific alterations in the brain.

4 Limitations

While the current study sheds new light onto the reliability of NV in comparison with RS, it comes with some limitations. Firstly, the reliability of graph metrics is strongly influenced by the choice of the applied preprocessing (Andellini et al., 2015). In this study, we applied motion correction and regressed out WM and CSF signals, as has been done in most previous studies (Braun et al., 2012; Cao et al., 2014; Wang et al., 2017). On top of that, we here applied basic (that is only removing the mean signal of the whole brain) global signal regression. There is an ongoing debate of whether or not to apply global signal regression, with some studies claiming that it introduces spurious anti-correlations while other reports suggest that these anti-correlations are true negative connections (Liang et al., 2012; Murphy et al., 2009). However, a review by Andellini et al., 2015 found no significant differences between the reliability of data with and without the inclusion of global signal regression across five studies (Andellini et al., 2015). Secondly, we here considered both, negative and positive connections, with the assumption that both are true representations of connectivity. However, several papers have indicated that negative correlations should be evaluated with care since they tend to reduce test-retest reliability (Andellini et al., 2015; Schwarz and McGonigle, 2011; Wang et al., 2011). Therefore, the reliability of single graph metrics in our study might have been decreased by the inclusion of negative connections. Thirdly, our results are based on weighted adjacency matrices, because they better characterize the underlying connectivity by considering connectivity strength. However, previous studies have suggested that binarized adjacency matrices may lead to higher reliability (Andellini et al., 2015; Wang et al., 2011). Nevertheless, we think that using weighted adjacency matrices is preferable, especially for clinical studies where subtle changes in connectivity might help to identify disease specific alterations.

Conclusion

NV has been suggested to improve the reliability of graph based measures in comparison with RS. Our findings extend the current knowledge by investigating this effect in different networks, with multiple NV stimuli and in two different cohorts. We demonstrate that the potential increase in reliability is dependent on the chosen NV stimuli and varies between functional networks. Furthermore we suggest that cultural differences should be considered when sharing NV stimuli across sites. Our study supports the use of NV to increase reliability of graph metrics, but emphasizes the need to carefully select the appropriate stimulus and network for the research question at hand.

Acknowledgment

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 945539 (HBP SGA3), and the Deutsche Forschungsgemeinschaft (491111487). We also acknowledge the funding support from Yong Loo Lin School of Medicine, National University of Singapore (J.H.Z), the Duke-NUS Medical School Signature Research Program Core Funding (J.H.Z.), and Ministry of Education, Singapore (MOE-T2EP40120-0007, J.H.Z), and Far East Organization (E-546-00-0398-01, MWLC.).

Conflict of interest statement

The authors declare that they have no conflict of interest.

A Hagberg, D Schult, P Swart, 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. Presented at the Proceedings of the 7th Python in Science conference (SciPy 2008).

Adams, R.B., Rule, N.O., Franklin, R.G., Wang, E., Stevenson, M.T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., Ambady, N., 2010. Cross-cultural Reading the Mind in the Eyes: An fMRI Investigation. Journal of Cognitive Neuroscience 22, 97–108. https://doi.org/10.1162/jocn.2009.21187

Alink, A., Schwiedrzik, C.M., Kohler, A., Singer, W., Muckli, L., 2010. Stimulus Predictability Reduces Responses in Primary Visual Cortex. Journal of Neuroscience 30, 2960–2966. https://doi.org/10.1523/JNEUROSCI.3730-10.2010

Amft, M., Bzdok, D., Laird, A.R., Fox, P.T., Schilbach, L., Eickhoff, S.B., 2015. Definition and characterization of an extended social-affective default network. Brain Struct Funct 220, 1031–1049. https://doi.org/10.1007/s00429-013-0698-0

Andellini, M., Cannatà, V., Gazzellini, S., Bernardi, B., Napolitano, A., 2015. Test-retest reliability of graph metrics of resting state MRI functional brain networks: A review. Journal of Neuroscience Methods 253, 183–192. https://doi.org/10.1016/j.jneumeth.2015.05.020

Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test–retest reliability of functional MRI in a classification learning task. NeuroImage 29, 1000–1006. https://doi.org/10.1016/j.neuroimage.2005.08.010

Avants, B., Tustison, N.J., Song, G., 2009. Advanced Normalization Tools: V1.0. The Insight Journal. https://doi.org/10.54294/uvnhin

Balthazar, M.L.F., de Campos, B.M., Franco, A.R., Damasceno, B.P., Cendes, F., 2014. Whole cortical and default mode network mean functional connectivity as potential biomarkers for mild Alzheimer's disease. Psychiatry Research: Neuroimaging 221, 37–42. https://doi.org/10.1016/j.pscychresns.2013.10.010

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., Filippi, M., 2019. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. NeuroImage: Clinical 21, 101645. https://doi.org/10.1016/j.nicl.2018.101645

Bassett, D.S., Bullmore, E., Verchinski, B.A., Mattay, V.S., Weinberger, D.R., Meyer-Lindenberg, A., 2008. Hierarchical Organization of Human Cortical Networks in Health and Schizophrenia. J. Neurosci. 28, 9239–9248. https://doi.org/10.1523/JNEUROSCI.1929-08.2008

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Annals of the New York Academy of Sciences 1191, 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. Cerebral Cortex 19, 2767–2796. https://doi.org/10.1093/cercor/bhp055

Braun, U., Plichta, M.M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., Walter, H., Seiferth, N., Kirsch, P., Meyer-Lindenberg, A., 2012. Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. NeuroImage 59, 1404–1412. https://doi.org/10.1016/j.neuroimage.2011.08.044

Buhle, J.T., Silvers, J.A., Wager, T.D., Lopez, R., Onyemekwu, C., Kober, H., Weber, J., Ochsner, K.N., 2014. Cognitive Reappraisal of Emotion: A Meta-Analysis of Human Neuroimaging Studies. Cerebral Cortex 24, 2981–2990. https://doi.org/10.1093/cercor/bht154

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10, 186–198. https://doi.org/10.1038/nrn2575

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A.R., Langner, R., Eickhoff, S.B., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct Funct 217, 783–796.

https://doi.org/10.1007/s00429-012-0380-y
Camilleri, J.A., Müller, V.I., Fox, P., Laird, A.R., Hoffstaedter, F., Kalenscher, T., Eickhoff, S.B., 2018. Definition and characterization of an extended multiple-demand network. NeuroImage 165, 138–147. https://doi.org/10.1016/j.neuroimage.2017.10.020
Cao, H., Plichta, M.M., Schäfer, A., Haddad, L., Grimm, O., Schneider, M., Esslinger, C., Kirsch, P., Meyer-Lindenberg, A., Tost, H., 2014. Test–retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. NeuroImage 84, 888–900. https://doi.org/10.1016/j.neuroimage.2013.09.013
Caspers, S., Zilles, K., Laird, A.R., Eickhoff, S.B., 2010. ALE meta-analysis of action observation and imitation in the human brain. NeuroImage 50, 1148–1167. https://doi.org/10.1016/j.neuroimage.2009.12.112
Christoff, K., Ream, J.M., Gabrieli, J.D.E., 2004. Neural Basis of Spontaneous thought Processes. Cortex 40, 623–630. https://doi.org/10.1016/S0010-9452(08)70158-8
Cieslik, E.C., Mueller, V.I., Eickhoff, C.R., Langner, R., Eickhoff, S.B., 2015. Three key regions for supervisory attentional control: evidence from neuroimaging meta-analyses. Neurosci Biobehav Rev 48, 22–34. https://doi.org/10.1016/j.neubiorev.2014.11.003
Deuker, L., Bullmore, E.T., Smith, M., Christensen, S., Nathan, P.J., Rockstroh, B., Bassett, D.S., 2009. Reproducibility of graph metrics of human brain functional networks. NeuroImage 47, 1460–1468. https://doi.org/10.1016/j.neuroimage.2009.05.035
DuPre, E., Hanke, M., Poline, J.-B., 2020. Nature abhors a paywall: How open science can realize the potential of naturalistic stimuli. NeuroImage 216, 116330. https://doi.org/10.1016/j.neuroimage.2019.116330
Eickhoff, S.B., Milham, M., Vanderwal, T., 2020. Towards clinical applications of movie fMRI. NeuroImage 217, 116860. https://doi.org/10.1016/j.neuroimage.2020.116860
Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S.S., Wright, J., Durnez, J., Poldrack, R.A., Gorgolewski, K.J., 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods 16, 111–116. https://doi.org/10.1038/s41592-018-0235-4
Finn, E.S., Bandettini, P.A., 2020. Movie-watching outperforms rest for functional connectivity-based prediction of behavior (preprint). Neuroscience. https://doi.org/10.1101/2020.08.23.263723
Finn, E.S., Scheinost, D., Finn, D.M., Shen, X., Papademetris, X., Constable, R.T., 2017. Can brain state be manipulated to emphasize individual differences in functional connectivity? NeuroImage 160, 140–151.
Gal, S., Coldham, Y., Tik, N., Bernstein-Eliav, M., Tavor, I., 2022. Act natural: Functional connectivity from naturalistic stimuli fMRI outperforms resting-state in predicting brain activity. NeuroImage 258, 119359. https://doi.org/10.1016/j.neuroimage.2022.119359
Goh, J.O.S., Leshikar, E.D., Sutton, B.P., Tan, J.C., Sim, S.K.Y., Hebrank, A.C., Park, D.C., 2010. Culture differences in neural processing of faces and houses in the ventral visual cortex. Soc Cogn Affect Neurosci 5, 227–235. https://doi.org/10.1093/scan/nsq060
Golland, Y., Bentin, S., Gelbard, H., Benjamini, Y., Heller, R., Nir, Y., Hasson, U., Malach, R., 2007. Extrinsic and Intrinsic Systems in the Posterior Cortex of the Human Brain Revealed during Natural Sensory Stimulation. Cerebral Cortex 17, 766–777. https://doi.org/10.1093/cercor/bhk030
Gonzalez-Castillo, J., Kam, J.W.Y., Hoy, C.W., Bandettini, P.A., 2021. How to Interpret Resting-State fMRI: Ask Your Participants. J. Neurosci. 41, 1130–1141. https://doi.org/10.1523/JNEUROSCI.1786-20.2020
Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. NeuroImage 48, 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060
Guo, C.C., Hyett, M.P., Nguyen, V.T., Parker, G.B., Breakspear, M.J., 2016. Distinct neurobiological signatures of brain connectivity in depression subtypes during natural

viewing of emotionally salient films. Psychol. Med. 46, 1535–1545. https://doi.org/10.1017/S0033291716000179

Guo, C.C., Kurth, F., Zhou, J., Mayer, E.A., Eickhoff, S.B., Kramer, J.H., Seeley, W.W., 2012. One-year test–retest reliability of intrinsic connectivity network fMRI in older adults. NeuroImage 61, 1471–1483. https://doi.org/10.1016/j.neuroimage.2012.03.027

Guo, C.C., Nguyen, V.T., Hyett, M.P., Parker, G.B., Breakspear, M.J., 2015. Out-of-sync: disrupted neural activity in emotional circuitry during film viewing in melancholic depression. Sci Rep 5, 11605. https://doi.org/10.1038/srep11605

Halchenko, Y., Meyer, K., Poldrack, B., Solanky, D., Wagner, A., Gors, J., MacFarlane, D., Pustina, D., Sochat, V., Ghosh, S., Mönch, C., Markiewicz, C., Waite, L., Shlyakhter, I., de la Vega, A., Hayashi, S., Häusler, C., Poline, J.-B., Kadelka, T., Skytén, K., Jarecka, D., Kennedy, D., Strauss, T., Cieslak, M., Vavra, P., Ioanas, H.-I., Schneider, R., Pflüger, M., Haxby, J., Eickhoff, S., Hanke, M., 2021. DataLad: distributed system for joint management of code, data, and their relationship. JOSS 6, 3262. https://doi.org/10.21105/joss.03262

Harada, T., Mano, Y., Komeda, H., Hechtman, L.A., Pornpattananangkul, N., Parrish, T.B., Sadato, N., Iidaka, T., Chiao, J.Y., 2020. Cultural influences on neural systems of intergroup emotion perception: An fMRI study. Neuropsychologia 137, 107254. https://doi.org/10.1016/j.neuropsychologia.2019.107254

Hasson, U., Furman, O., Clark, D., Dudai, Y., Davachi, L., 2008. Enhanced Intersubject Correlations during Movie Viewing Correlate with Successful Episodic Encoding. Neuron 57, 452–462. https://doi.org/10.1016/j.neuron.2007.12.009

Hasson, U., Malach, R., Heeger, D.J., 2010. Reliability of cortical activity during natural stimulation. Trends in Cognitive Sciences 14, 40–48. https://doi.org/10.1016/j.tics.2009.10.011

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject Synchronization of Cortical Activity During Natural Vision. Science 303, 1634–1640. https://doi.org/10.1126/science.1089506

Hlinka, J., Děchtěrenko, F., Rydlo, J., Androvičová, R., Vejmelka, M., Jajcay, L., Tintěra, J., Lukavský, J., Horáček, J., 2022. The intra-session reliability of functional connectivity during naturalistic viewing conditions. Psychophysiology 59, e14075. https://doi.org/10.1111/psyp.14075

Hyett, M.P., Parker, G.B., Guo, C.C., Zalesky, A., Nguyen, V.T., Yuen, T., Breakspear, M., 2015. Scene unseen: Disrupted neuronal adaptation in melancholia during emotional film viewing. NeuroImage: Clinical 9, 660–667. https://doi.org/10.1016/j.nicl.2015.10.011

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage 17, 825–841. https://doi.org/10.1006/nimg.2002.1132

Ki, J.J., Kelly, S.P., Parra, L.C., 2016. Attention Strongly Modulates Reliability of Neural Responses to Naturalistic Narrative Stimuli. J. Neurosci. 36, 3092–3101. https://doi.org/10.1523/JNEUROSCI.2942-15.2016

Kong, J., Gollub, R.L., Webb, J.M., Kong, J.-T., Vangel, M.G., Kwong, K., 2007. Test–retest study of fMRI signal change evoked by electroacupuncture stimulation. NeuroImage 34, 1171–1181. https://doi.org/10.1016/j.neuroimage.2006.10.019

Koster-Hale, J., Saxe, R., 2013. Theory of Mind: A Neural Prediction Problem. Neuron 79, 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kröll, J.-P., Friedrich, P., Li, X., Patil, K.R., Mochalski, L., Waite, L., Qian, X., Chee, M.W., Zhou, J.H., Eickhoff, S., Weis, S., 2023. Naturalistic viewing increases individual identifiability based on connectivity within functional brain networks. NeuroImage 120083. https://doi.org/10.1016/j.neuroimage.2023.120083

Langner, R., Eickhoff, S.B., 2013. Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. Psychological Bulletin 139, 870–900. https://doi.org/10.1037/a0030694

Li, X., Friedrich, P., Patil, K.R., Eickhoff, S.B., Weis, S., 2022. A topography-based predictive framework for naturalistic viewing fMRI (preprint). Neuroscience. https://doi.org/10.1101/2022.05.26.493420

Liang, X., Wang, J., Yan, C., Shu, N., Xu, K., Gong, G., He, Y., 2012. Effects of Different Correlation Metrics and Preprocessing Factors on Small-World Brain Functional Networks: A Resting-State Functional MRI Study. PLoS ONE 7, e32766. https://doi.org/10.1371/journal.pone.0032766

Liu, X., Hairston, J., Schrier, M., Fan, J., 2011. Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. Neuroscience & Biobehavioral Reviews 35, 1219–1236. https://doi.org/10.1016/j.neubiorev.2010.12.012

Mason, M.F., Norton, M.I., Van Horn, J.D., Wegner, D.M., Grafton, S.T., Macrae, C.N., 2007. Wandering Minds: The Default Network and Stimulus-Independent Thought. Science 315, 393–395. https://doi.org/10.1126/science.1131295

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Müller, R., Büttner, P., 1994. A critical discussion of intraclass correlation coefficients. Statistics in Medicine 13, 2465–2476. https://doi.org/10.1002/sim.4780132310

Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? NeuroImage 44, 893–905. https://doi.org/10.1016/j.neuroimage.2008.09.036

O'Connor, D., Potler, N.V., Kovacs, M., Xu, T., Ai, L., Pellman, J., Vanderwal, T., Parra, L.C., Cohen, S., Ghosh, S., Escalera, J., Grant-Villegas, N., Osman, Y., Bui, A., Craddock, R.C., Milham, M.P., 2017. The Healthy Brain Network Serial Scanning Initiative: a resource for evaluating inter-individual differences and their reliabilities across scan conditions and sessions. GigaScience 6. https://doi.org/10.1093/gigascience/giw011

Petrella, J.R., 2011. Use of Graph Theory to Evaluate Brain Networks: A Clinical Tool for a Small World? Radiology 259, 317–320. https://doi.org/10.1148/radiol.11110380

Reijneveld, J.C., Ponten, S.C., Berendse, H.W., Stam, C.J., 2007. The application of graph theoretical analysis to complex networks in the brain. Clinical Neurophysiology 118, 2317–2331. https://doi.org/10.1016/j.clinph.2007.08.010

Rikandi, E., Mäntylä, T., Lindgren, M., Kieseppä, T., Suvisaari, J., Raij, T.T., 2022. Functional network connectivity and topology during naturalistic stimulus is altered in first-episode psychosis. Schizophrenia Research 241, 83–91. https://doi.org/10.1016/j.schres.2022.01.006

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A.R., Schulz, J.B., Fox, P.T., Eickhoff, S.B., 2012. Modelling neural correlates of working memory: A coordinate-based meta-analysis. NeuroImage 60, 830–846. https://doi.org/10.1016/j.neuroimage.2011.11.050

Rubinov, M., Knock, S.A., Stam, C.J., Micheloyannis, S., Harris, A.W.F., Williams, L.M., Breakspear, M., 2009. Small-world properties of nonlinear brain activity in schizophrenia. Hum. Brain Mapp. 30, 403–416. https://doi.org/10.1002/hbm.20517

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: Uses and interpretations. NeuroImage 52, 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003

Saarimäki, H., 2021. Naturalistic Stimuli in Affective Neuroimaging: A Review. Front. Hum. Neurosci. 15, 675068. https://doi.org/10.3389/fnhum.2021.675068

Sabatinelli, D., Fortune, E.E., Li, Q., Siddiqui, A., Krafft, C., Oliver, W.T., Beck, S., Jeffries, J., 2011. Emotional perception: Meta-analyses of face and natural scene processing. NeuroImage 54, 2524–2533. https://doi.org/10.1016/j.neuroimage.2010.10.011

Schaefer, A., Nils, F., Sanchez, X., Philippot, P., 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cognition & Emotion 24, 1153–1172. https://doi.org/10.1080/02699930903274322

Schwarz, A.J., McGonigle, J., 2011. Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. NeuroImage 55, 1132–1146. https://doi.org/10.1016/j.neuroimage.2010.12.047

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sneddon, I., McKeown, G., McRorie, M., Vukicevic, T., 2011. Cross-Cultural Patterns in Dynamic Ratings of Positive and Negative Natural Emotional Behaviour. PLoS ONE 6, e14679. https://doi.org/10.1371/journal.pone.0014679

Spreng, R.N., Mar, R.A., Kim, A.S.N., 2009. The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind, and the Default Mode: A Quantitative Meta-analysis. Journal of Cognitive Neuroscience 21, 489–510. https://doi.org/10.1162/jocn.2008.21029

Stam, C.J., Reijneveld, J.C., 2007. Graph theoretical analysis of complex networks in the brain. Nonlinear Biomed Phys 1, 3. https://doi.org/10.1186/1753-4631-1-3

Supekar, K., Menon, V., Rubin, D., Musen, M., Greicius, M.D., 2008. Network Analysis of Intrinsic Functional Brain Connectivity in Alzheimer's Disease. PLoS Comput Biol 4, e1000100. https://doi.org/10.1371/journal.pcbi.1000100

Synchon Mandal, Raimondo, F., Sasse, L., Komeyer, V., Kaustubh Patil, Hamdan, S., Hoffstaedter, F., Poldrack, B., Weiss, S., 2023. juaml/junifer: v0.0.3. https://doi.org/10.5281/ZENODO.8176569

Tagliazucchi, E., Laufs, H., 2014. Decoding Wakefulness Levels from Typical fMRI Resting-State Data Reveals Reliable Drifts between Wakefulness and Sleep. Neuron 82, 695–708. https://doi.org/10.1016/j.neuron.2014.03.020

Telesford, Q.K., Morgan, A.R., Hayasaka, S., Simpson, S.L., Barret, W., Kraft, R.A., Mozolic, J.L., Laurienti, P.J., 2010. Reproducibility of Graph Metrics in fMRI Networks. Front. Neuroinform. 4. https://doi.org/10.3389/fninf.2010.00117

Tian, L., Ye, M., Chen, C., Cao, X., Shen, T., 2021. Consistency of functional connectivity across different movies. NeuroImage 233, 117926. https://doi.org/10.1016/j.neuroimage.2021.117926

Van Dijk, K.R.A., Sabuncu, M.R., Buckner, R.L., 2012. The influence of head motion on intrinsic functional connectivity MRI. NeuroImage 59, 431–438. https://doi.org/10.1016/j.neuroimage.2011.07.044

Vanderwal, T., Eilbott, J., Finn, E.S., Craddock, R.C., Turnbull, A., Castellanos, F.X., 2017. Individual differences in functional connectivity during naturalistic viewing conditions. NeuroImage 157, 521–530. https://doi.org/10.1016/j.neuroimage.2017.06.027

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L.C., Castellanos, F.X., 2015. Inscapes : A movie paradigm to improve compliance in functional magnetic resonance imaging. NeuroImage 122, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069

Wang, J., Ren, Y., Hu, X., Nguyen, V.T., Guo, L., Han, J., Guo, C.C., 2017. Test-retest reliability of functional connectivity networks during naturalistic fMRI paradigms: Test-Retest Reliability of Naturalistic fMRI. Hum. Brain Mapp. 38, 2226–2241. https://doi.org/10.1002/hbm.23517

Wang, J.-H., Zuo, X.-N., Gohel, S., Milham, M.P., Biswal, B.B., He, Y., 2011. Graph Theoretical Analysis of Functional Brain Networks: Test-Retest Evaluation on Short- and Long-Term Resting-State Functional MRI Data. PLoS ONE 6, e21976. https://doi.org/10.1371/journal.pone.0021976

Wang, L., Zhu, C., He, Y., Zang, Y., Cao, Q., Zhang, H., Zhong, Q., Wang, Y., 2009. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. Hum. Brain Mapp. 30, 638–649. https://doi.org/10.1002/hbm.20530

Witt, S.T., Laird, A.R., Meyerand, M.E., 2008. Functional neuroimaging correlates of finger-tapping task variations: An ALE meta-analysis. NeuroImage 42, 343–356. https://doi.org/10.1016/j.neuroimage.2008.04.025

Wu, T., Wang, L., Chen, Y., Zhao, C., Li, K., Chan, P., 2009. Changes of functional connectivity of the motor network in the resting state in Parkinson's disease. Neuroscience Letters 460, 6–10. https://doi.org/10.1016/j.neulet.2009.05.046

Yang, Z., Wu, J., Xu, L., Deng, Z., Tang, Y., Gao, J., Hu, Y., Zhang, Y., Qin, S., Li, C., Wang, J., 2020. Individualized psychiatric imaging based on inter-subject neural synchronization in movie watching. NeuroImage 216, 116227. https://doi.org/10.1016/j.neuroimage.2019.116227

Zhang, G., Liu, X., 2021. Investigation of functional brain network reconfiguration during exposure to naturalistic stimuli using graph-theoretical analysis. J. Neural Eng. 18, 056027. https://doi.org/10.1088/1741-2552/ac20e7

Zhang, H., Zhang, Y.-J., Duan, L., Ma, S.-Y., Lu, C.-M., Zhu, C.-Z., 2011. Is resting-state functional connectivity revealed by functional near-infrared spectroscopy test-retest reliable? J. Biomed. Opt. 16, 067008. https://doi.org/10.1117/1.3591020

Zhang, X., Liu, J., Yang, Y., Zhao, S., Guo, L., Han, J., Hu, X., 2022. Test–retest reliability of dynamic functional connectivity in naturalistic paradigm functional magnetic resonance imaging. Human Brain Mapping 43, 1463–1476. https://doi.org/10.1002/hbm.25736

5 Discussion

5.1 Interpretable ML frameworks

Despite the opportunities provided by large datasets and ever more advanced ML algorithms, only a fraction of proposed ML methods make it to clinical application. One of the major hurdles for clinical translation is the interpretability of a given method (Dinsdale et al., 2022; Thibeau-Sutre et al., 2023). With first disease-modifying treatments becoming available for AD (Mintun et al., 2021; Van Dyck et al., 2023), accurate diagnosis and prognosis of the disease have become even more urgent. Likewise, this places a high demand on the explainability of potential biomarkers as decisions based on these biomarkers will lead to drug-administration.Various examples have shown that "blackbox" models applied in clinical settings can produce seemingly accurate results, relying on confounders, but will fail to generalize on new data (DeGrave et al., 2021; Thibeau-Sutre et al., 2023; Winkler et al., 2019). This dissertation provides a framework that not only improves accurate diagnosis and prognosis of AD, by constructing complex features, but also maintains interpretability of the constructed features. For both applications, the features constructed by the GE framework improved performance across all four metrics used in the study, in comparison with models using base features (Study 1, Table 3, 4). The performance of the models were comparable to results reported in other studies that employed explainable ML frameworks for diagnosis and prognosis of AD (Bloch et al., 2021; Bogdanovic et al., 2022; Böhle et al., 2019; Pohl et al., 2022). The features that were constructed by the GE framework integrate information about the complex interactions between base features, such that the result is still interpretable. An analysis of the constructed features showed that they combined brain regions that are known to be affected in AD, such as the temporal pole (Scheltens et al., 1992), Amygdala (Poulin et al., 2011), Putamen (de Jong et al., 2008) and Thalamus (de Jong et al., 2008). More so, the constructed features were still interpretable as it was observable that they contained information about the atrophy or co-atrophy of AD-involved brain regions.

5.2 Individual differences during RS and NV

One of the primary goals of neuroscience is to relate differences in brain functions to differences in phenotypes. However, analyzing differences in individual FC patterns that occur during RS has led to unsatisfactory results. Therefore, study 2 investigates how a new paradigm,

NV, enhances individual differences in comparison with RS. By calculating the identifiability of individual FC matrices extracted during three NV stimuli and RS, this dissertation provides clear evidence for improved detection of individual differences during movie-watching. Using identifiability or "fingerprinting" of FC matrices as a proxy for individual differences has been made popular by Finn et al (Finn et al., 2015) and was previously used by Vanderwal et al to show that NV can enhance individual differences on a whole brain basis (Vanderwal et al., 2017). Similarly, identifiability was used in this dissertation to compare the effects of NV and RS on individual differences in functional networks. In ten out of fourteen networks, NV improved identification accuracy over RS (Study 2, Tab.1). The improvement seen for NV was most prominent for the Indiana Jones stimulus, which led to the highest identifiability in eight of the networks. On the other hand, the movie Inscapes was generally similar or inferior to RS, while Circus showed improvement in only two networks. These results are in line with previous studies that have suggested that in order to maximally engage the participant, NV stimuli with more social content might be preferable over neutral/abstract stimuli (Dmochowski et al., 2014; Finn and Bandettini, 2020; Nummenmaa et al., 2014; Schmälzle et al., 2015). This notion was further supported by comparing patterns of inter-individual NFC between conditions. NFC patterns during Inscapes were mostly similar to those during RS, while Circus and Indiana Jones exhibit connectivity profiles that are distinct from RS across networks (Study 2, Fig.1, Fig. 2).

5.3 Variability across functional networks

The vast majority of studies that investigate the impact of NV have focused on whole-brain connectivity. This dissertation instead focuses on connectivity on a network level and provides evidence that the effect of NV deviates across networks covering different cognitive domains. Study 2 investigated three different NV stimuli and revealed differences in within- and between-subject correlations during RS and NV that were obscured on a whole-brain level (Study 2, Fig 4, 5, S1, S2). Based on the overall increased identifiability during NV in Study 2, one might expect that within-subject correlations (as a measure for stable individual patterns) are increased during NV as well. However, identifiability is always dependent on the ratio of within- and between-subject correlations, e.g. subjects that are too similar to each other will be harder to identify even though they might exhibit stable patterns across sessions (Finn et al., 2017). In Study 2, increased within-subject correlations were observed in meta-analytic networks that are essential for the perception and processing of action, behavior and emotions. With regards to the assumption that the social aspect of a movie stimulus induces stable individual connectivity patterns, it is only reasonable to expect that this effect is more

pronounced in networks that deal with the processing of social interactions. On a whole-brain level, between-subject correlations are generally presumed to be increased by NV, given that all subjects are presented with the same stimulus (Hasson et al., 2004; Vanderwal et al., 2017). However, this dissertation importantly shows that this effect is not unambiguously true across functional networks. NV increased between-subject correlations in networks that are associated with executive functions and/or stimulus evaluation. On the other hand, networks that are more related to intrinsically oriented functions exhibited reduced between-subject correlations during NV. Presumably, the function of these networks is suppressed during the processing of complex stimuli, thus preventing coordinated activity in these networks which in turn reduces similarity between subjects.

This dissertation also investigated within- and between-subject correlations in functional networks during a full narrative movie (FNM), Forrest Gump, from studyforrest project (Hanke et al., 2016). Contrary to the shorter stimuli used in Study 2 (10 minutes), a FNM provides emotions embedded in a richer context and evolving over a longer time, allowing for a more comprehensive study of socio-affective processes. Study 3 showed that the effect of the FNM on changes in within- and between-subject correlations was dependent on the network (Study 3, Fig. 3), confirming results from Study 2. Furthermore, Study 3 implemented linear mixed models to analyze how the narrative of the movie and the portrayed valence and arousal affected within- and between-subject correlations across networks. Based on valence and arousal annotations from Labs et al (Labs et al., 2015), the analysis revealed that within- and between-subject correlations were best accounted for by network, movie segment, valence and a movie segment by valence interaction. Within-subject correlations were further explained by an interaction of movie segment, valence and arousal. Taken together, these findings show that within- and between-subject correlations during NV are sensitive to the progressing narrative and emotions portrayed in a stimulus and differ between networks. Lower within-subject correlations during the FNM were observed in the AM, ER, SM, ToM, and eSAD networks (Study 3, Fig. 3), which align with patterns observed in Study 2 that show a tendency for lower within-subject correlations during NV than during RS, in these networks (Study 2, Fig. 4). In addition, Study 3 could also demonstrate that within- and between-subject correlations across networks increased as the movie progressed, suggesting a general trend towards greater similarity in subjects' NFC over time (Study 3, Fig. 3). Related, previous work has demonstrated that certain cognitive and emotional processes develop only over extended time periods (Hasson et al., 2010). This dissertation shows that both, movie clips and FNMs, have different effects

across functional networks, emphasizing that a network perspective grants more detailed insights into the full effect of NV paradigms than whole-brain analysis.

5.4 Reliability of NV stimuli

For any research question at hand, the crucial prerequisite is that the used measurement is reliable, such that differences across subjects and time points can be meaningfully interpreted. This dissertation investigates the reliability of NV paradigms and compares it to that of conventional RS. Study 4 shows that NV can improve reliability over RS across networks dealing with affective, social, executive, memory and motor functions, in two samples (Study 4, Fig. 1, Fig. 2). These results indicate that NV can increase engagement not only in sensory, but also in higher order networks. The observed reliability in study 4 matched results from previous studies investigating graph metrics extracted from RS and NV (Braun et al., 2012; Cao et al., 2014; Wang et al., 2017). Similar to Wang et al (Wang et al., 2017) and results from study 2 and study 3, effects of NV varied across networks. However, in contrast to results from Wang et al, where the majority of networks showed improved reliability during NV, NV was less reliable than RS in the majority of networks and graph metrics in study 4. Possibly, since sessions for both datasets were conducted within a week, participants might have been rather familiar with the movie stimuli during the second session. Multiple studies have shown that expected stimuli reduce the neuronal response (Alink et al., 2010; Koster-Hale and Saxe, 2013), which in turn might have led to the relative decrease in reliability for NV here. Still, NV at least partially increased reliability over RS. In these cases, it was again observable that the NV stimuli with more social content, Circus and Indiana Jones, improved reliability more often than Inscapes. In addition, Study 4 investigated if the effect of NV stimuli is different across cohorts with different cultural backgrounds. Previous studies have demonstrated cultural differences for the perception of faces (Adams et al., 2010; Goh et al., 2010; Harada et al., 2020) and rating of emotions when watching movie clips (Sneddon et al., 2011). Similarly, the results of the Asian and European cohort in Study 4 were mostly different across stimulus, network and graph metric (Study 4, Fig. 3). Therefore, future studies should consider the cultural background of a cohort when choosing a movie stimulus.

5.5 Conclusions

This dissertation addressed primary challenges for the translation of MRI based biomarkers into clinical use, such as accuracy, reliability and interpretability. Therefore, a simple GE based

framework was provided that constructs complex feature representations while remaining interpretability. The GE framework was demonstrated to be applicable to the diagnosis and prognosis of AD, one of the most prevalent neurological diseases as of today, where it could significantly improve predictive performance. Subsequent inspection of the features uncovered humanly interpretable patterns of co-atrophy in brain regions typically impacted by AD. Further, this dissertation investigated if NV paradigms can improve key biomarker metrics such as reliability, reduced intra-subject variability and enhanced detection of individual differences, in comparison with RS. Therefore, different NV stimuli with varying levels of social content, as well as different lengths and their effect in functional brain networks were compared. The comparison of different NV stimuli revealed that certain stimuli, The Circus and Indiana Jones, are better suited to improve the detection of individual differences, possibly due to a higher level of social content. A clustering of the connectivity profiles during the different stimuli confirmed that these two stimuli were more distinct from RS than the movie Inscapes, which lacks social interaction. Further, an analysis of within- and between subject correlations demonstrated that shorter movie clips as well as a FNM can improve similarity within and between subjects, in comparison with RS. In addition, it was shown that NV stimuli can increase the reliability of fMRI, as measured by graph metrics. This dissertation extends the current knowledge about NV paradigms by examining their effect in functional networks. Contrary to previous studies that focused on whole-brain, it was demonstrated that NV stimuli do not unconditionally improve reliability, as well as within- and between-subject correlations across the brain, but rather that the effect varies between functional networks. Especially in networks that are related to intrinsically related functions, RS was shown to be preferable over NV.

Looking forward, the provided GE framework can be helpful in future biomarker studies where interpretability of a model is a must, by promoting both accuracy and interpretability. As drug development for neurological diseases advances, biomarkers that diagnose and monitor these diseases will become increasingly important, and methods like the proposed framework have the potential to play a crucial role in the development of such biomarkers. Further, the results here encourage the use of NV stimuli to improve signal properties of fMRI, that are important for biomarker research. However, the results highlight the importance to carefully chose the appropriate stimulus for the research question at hand. Generally, NV stimuli with social content should be preferred. Future biomarker studies might benefit from NV paradigms by selecting a stimulus that is specific to their research focus such as anxiety-inducing movie clips to probe patients with anxiety or a NV stimulus with distractors for patients with attention-

deficit/hyperactivity disorder (ADHD). Finally, this dissertation provides a new NV dataset, which employs three NV stimuli with different levels of social content, that is publicly available to the neuroimaging community.

6 References

Adams, R.B., Rule, N.O., Franklin, R.G., Wang, E., Stevenson, M.T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., Ambady, N., 2010. Cross-cultural Reading the Mind in the Eyes: An fMRI Investigation. Journal of Cognitive Neuroscience 22, 97–108. https://doi.org/10.1162/jocn.2009.21187

Alink, A., Schwiedrzik, C.M., Kohler, A., Singer, W., Muckli, L., 2010. Stimulus Predictability Reduces Responses in Primary Visual Cortex. Journal of Neuroscience 30, 2960–2966. https://doi.org/10.1523/JNEUROSCI.3730-10.2010

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Annals of the New York Academy of Sciences 1191, 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x

Bloch, L., Friedrich, C.M., for the Alzheimer's Disease Neuroimaging Initiative, 2021. Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning. Alz Res Therapy 13, 155. https://doi.org/10.1186/s13195-021-00879-4

Bogdanovic, B., Eftimov, T., Simjanoska, M., 2022. In-depth insights into Alzheimer's disease by using explainable machine learning approach. Sci Rep 12, 6508. https://doi.org/10.1038/s41598-022-10202-2

Böhle, M., Eitel, F., Weygandt, M., Ritter, K., 2019. Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. Front. Aging Neurosci. 11, 194. https://doi.org/10.3389/fnagi.2019.00194

Braun, U., Plichta, M.M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., Walter, H., Seiferth, N., Kirsch, P., Meyer-Lindenberg, A., 2012. Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. NeuroImage 59, 1404–1412. https://doi.org/10.1016/j.neuroimage.2011.08.044

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A.R., Langner, R., Eickhoff, S.B., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct Funct 217, 783–796. https://doi.org/10.1007/s00429-012-0380-y

Cao, H., Plichta, M.M., Schäfer, A., Haddad, L., Grimm, O., Schneider, M., Esslinger, C., Kirsch, P., Meyer-Lindenberg, A., Tost, H., 2014. Test–retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. NeuroImage 84, 888–900. https://doi.org/10.1016/j.neuroimage.2013.09.013

Christoff, K., Ream, J.M., Gabrieli, J.D.E., 2004. Neural Basis of Spontaneous thought Processes.

Cortex 40, 623–630. https://doi.org/10.1016/S0010-9452(08)70158-8

de Jong, L.W., van der Hiele, K., Veer, I.M., Houwing, J.J., Westendorp, R.G.J., Bollen, E.L.E.M., de Bruin, P.W., Middelkoop, H.A.M., van Buchem, M.A., van der Grond, J., 2008. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. Brain 131, 3277–3285. https://doi.org/10.1093/brain/awn278

DeGrave, A.J., Janizek, J.D., Lee, S.-I., 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 3, 610–619. https://doi.org/10.1038/s42256-021-00338-7

Dinsdale, N.K., Bluemke, E., Sundaresan, V., Jenkinson, M., Smith, S.M., Namburete, A.I.L., 2022. Challenges for machine learning in clinical translation of big data imaging studies. Neuron 110, 3866–3881. https://doi.org/10.1016/j.neuron.2022.09.012

Dmochowski, J.P., Bezdek, M.A., Abelson, B.P., Johnson, J.S., Schumacher, E.H., Parra, L.C., 2014. Audience preferences are predicted by temporal reliability of neural processing. Nat Commun 5, 4567. https://doi.org/10.1038/ncomms5567

Dubois, J., Adolphs, R., 2016. Building a Science of Individual Differences from fMRI. Trends in Cognitive Sciences 20, 425–443. https://doi.org/10.1016/j.tics.2016.03.014

DuPre, E., Hanke, M., Poline, J.-B., 2020. Nature abhors a paywall: How open science can realize the potential of naturalistic stimuli. NeuroImage 216, 116330. https://doi.org/10.1016/j.neuroimage.2019.116330

Eickhoff, S.B., Bzdok, D., Laird, A.R., Kurth, F., Fox, P.T., 2012. Activation likelihood estimation meta-analysis revisited. NeuroImage 59, 2349–2361. https://doi.org/10.1016/j.neuroimage.2011.09.017

Eickhoff, S.B., Milham, M., Vanderwal, T., 2020. Towards clinical applications of movie fMRI. NeuroImage 217, 116860. https://doi.org/10.1016/j.neuroimage.2020.116860

EU guidelines on ethics in artificial intelligence: Context and implementation, 2019.

Finn, E.S., Bandettini, P.A., 2020. Movie-watching outperforms rest for functional connectivity-based prediction of behavior (preprint). Neuroscience. https://doi.org/10.1101/2020.08.23.263723

Finn, E.S., Corlett, P.R., Chen, G., Bandettini, P.A., Constable, R.T., 2018. Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. Nat Commun 9, 2043. https://doi.org/10.1038/s41467-018-04387-2

Finn, E.S., Scheinost, D., Finn, D.M., Shen, X., Papademetris, X., Constable, R.T., 2017. Can brain state be manipulated to emphasize individual differences in functional connectivity? NeuroImage 160, 140–151.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat Neurosci 18, 1664–1671. https://doi.org/10.1038/nn.4135

Goh, J.O.S., Leshikar, E.D., Sutton, B.P., Tan, J.C., Sim, S.K.Y., Hebrank, A.C., Park, D.C., 2010.

Culture differences in neural processing of faces and houses in the ventral visual cortex. Soc Cogn Affect Neurosci 5, 227–235. https://doi.org/10.1093/scan/nsq060

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F.R., Nigbur, R., Waite, A.Q., Baumgartner, F., Stadler, J., 2016. A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. Sci Data 3, 160092. https://doi.org/10.1038/sdata.2016.92

Harada, T., Mano, Y., Komeda, H., Hechtman, L.A., Pornpattananangkul, N., Parrish, T.B., Sadato, N., Iidaka, T., Chiao, J.Y., 2020. Cultural influences on neural systems of intergroup emotion perception: An fMRI study. Neuropsychologia 137, 107254. https://doi.org/10.1016/j.neuropsychologia.2019.107254

Hasson, U., Malach, R., Heeger, D.J., 2010. Reliability of cortical activity during natural stimulation. Trends in Cognitive Sciences 14, 40–48. https://doi.org/10.1016/j.tics.2009.10.011

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject Synchronization of Cortical Activity During Natural Vision. Science 303, 1634–1640. https://doi.org/10.1126/science.1089506

Koster-Hale, J., Saxe, R., 2013. Theory of Mind: A Neural Prediction Problem. Neuron 79, 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Labs, A., Reich, T., Schulenburg, H., Boennen, M., Mareike, G., Golz, M., Hartigs, B., Hoffmann, N., Keil, S., Perlow, M., Peukmann, A.K., Rabe, L.N., Von Sobbe, F.-R., Hanke, M., 2015. Portrayed emotions in the movie "Forrest Gump." F1000Res 4, 92. https://doi.org/10.12688/f1000research.6230.1

Lahmiri, S., Shmuel, A., 2019. Performance of machine learning methods applied to structural MRI and ADAS cognitive scores in diagnosing Alzheimer's disease. Biomedical Signal Processing and Control 52, 414–419. https://doi.org/10.1016/j.bspc.2018.08.009

Mintun, M.A., Lo, A.C., Duggan Evans, C., Wessels, A.M., Ardayfio, P.A., Andersen, S.W., Shcherbinin, S., Sparks, J., Sims, J.R., Brys, M., Apostolova, L.G., Salloway, S.P., Skovronsky, D.M., 2021. Donanemab in Early Alzheimer's Disease. N Engl J Med 384, 1691–1704. https://doi.org/10.1056/NEJMoa2100708

Nandi, A., Counts, N., Chen, S., Seligman, B., Tortorice, D., Vigo, D., Bloom, D.E., 2022. Global and regional projections of the economic burden of Alzheimer's disease and related dementias from 2019 to 2050: A value of statistical life approach. eClinicalMedicine 51, 101580. https://doi.org/10.1016/j.eclinm.2022.101580

Nguyen, M., Vanderwal, T., Hasson, U., 2019. Shared understanding of narratives is correlated with shared neural responses. NeuroImage 184, 161–170. https://doi.org/10.1016/j.neuroimage.2018.09.010

Nummenmaa, L., Saarimäki, H., Glerean, E., Gotsopoulos, A., Jääskeläinen, I.P., Hari, R., Sams, M., 2014. Emotional speech synchronizes brains across listeners and engages large-scale dynamic

brain networks. NeuroImage 102, 498–509. https://doi.org/10.1016/j.neuroimage.2014.07.063

Pohl, T., Jakab, M., Benesova, W., 2022. Interpretability of deep neural networks used for the diagnosis of Alzheimer's disease. Int J Imaging Syst Tech 32, 673–686. https://doi.org/10.1002/ima.22657

Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., 2011. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Research: Neuroimaging 194, 7–13. https://doi.org/10.1016/j.pscychresns.2011.06.014

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., Petersen, S.E., 2011. Functional Network Organization of the Human Brain. Neuron 72, 665–678. https://doi.org/10.1016/j.neuron.2011.09.006

Rikandi, E., Pamilo, S., Mäntylä, T., Suvisaari, J., Kieseppä, T., Hari, R., Seppä, M., Raij, T.T., 2017. Precuneus functioning differentiates first-episode psychosis patients during the fantasy movie Alice in Wonderland. Psychol. Med. 47, 495–506. https://doi.org/10.1017/S0033291716002609

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A.R., Schulz, J.B., Fox, P.T., Eickhoff, S.B., 2012. Modelling neural correlates of working memory: A coordinate-based meta-analysis. NeuroImage 60, 830–846. https://doi.org/10.1016/j.neuroimage.2011.11.050

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex 28, 3095–3114. https://doi.org/10.1093/cercor/bhx179

Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H.C., Vermersch, P., Kuiper, M., Steinling, M., Wolters, E.C., Valk, J., 1992. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. Journal of Neurology, Neurosurgery & Psychiatry 55, 967–972. https://doi.org/10.1136/jnnp.55.10.967

Schmälzle, R., Häcker, F.E.K., Honey, C.J., Hasson, U., 2015. Engaged listeners: shared neural processing of powerful political speeches. Social Cognitive and Affective Neuroscience 10, 1137–1143. https://doi.org/10.1093/scan/nsu168

Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R., Beckmann, C.F., 2009. Correspondence of the brain's functional architecture during activation and rest. Proc. Natl. Acad. Sci. U.S.A. 106, 13040–13045. https://doi.org/10.1073/pnas.0905267106

Sneddon, I., McKeown, G., McRorie, M., Vukicevic, T., 2011. Cross-Cultural Patterns in Dynamic Ratings of Positive and Negative Natural Emotional Behaviour. PLoS ONE 6, e14679. https://doi.org/10.1371/journal.pone.0014679

Sporns, O., Betzel, R.F., 2016. Modular Brain Networks. Annu. Rev. Psychol. 67, 613–640. https://doi.org/10.1146/annurev-psych-122414-033634

Spreng, R.N., Mar, R.A., Kim, A.S.N., 2009. The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind, and the Default Mode: A Quantitative Meta-analysis. Journal of Cognitive Neuroscience 21, 489–510. https://doi.org/10.1162/jocn.2008.21029

Thibeau-Sutre, E., Collin, S., Burgos, N., Colliot, O., 2023. Interpretability of Machine Learning Methods Applied to Neuroimaging, in: Colliot, O. (Ed.), Machine Learning for Brain Disorders, Neuromethods. Springer US, New York, NY, pp. 655–704. https://doi.org/10.1007/978-1-0716-3195-9_22

Van Dyck, C.H., Swanson, C.J., Aisen, P., Bateman, R.J., Chen, C., Gee, M., Kanekiyo, M., Li, D., Reyderman, L., Cohen, S., Froelich, L., Katayama, S., Sabbagh, M., Vellas, B., Watson, D., Dhadda, S., Irizarry, M., Kramer, L.D., Iwatsubo, T., 2023. Lecanemab in Early Alzheimer's Disease. N Engl J Med 388, 9–21. https://doi.org/10.1056/NEJMoa2212948

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Della Penna, S., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., 2012. The Human Connectome Project: A data acquisition perspective. NeuroImage 62, 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Vanderwal, T., Eilbott, J., Finn, E.S., Craddock, R.C., Turnbull, A., Castellanos, F.X., 2017. Individual differences in functional connectivity during naturalistic viewing conditions. NeuroImage 157, 521–530. https://doi.org/10.1016/j.neuroimage.2017.06.027

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L.C., Castellanos, F.X., 2015. Inscapes : A movie paradigm to improve compliance in functional magnetic resonance imaging. NeuroImage 122, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069

Wang, J., Ren, Y., Hu, X., Nguyen, V.T., Guo, L., Han, J., Guo, C.C., 2017. Test-retest reliability of functional connectivity networks during naturalistic fMRI paradigms: Test-Retest Reliability of Naturalistic fMRI. Hum. Brain Mapp. 38, 2226–2241. https://doi.org/10.1002/hbm.23517

Winkler, J.K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., Haenssle, H.A., 2019. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. JAMA Dermatol 155, 1135. https://doi.org/10.1001/jamadermatol.2019.1735

Zhu, W., Sun, L., Huang, J., Han, L., Zhang, D., 2021. Dual Attention Multi-Instance Deep Learning for Alzheimer's Disease Diagnosis With Structural MRI. IEEE Trans. Med. Imaging 40, 2354–2366. https://doi.org/10.1109/TMI.2021.3077079

Acknowledgements

I would first like to thank Prof. Dr. Simon Eickhoff for the opportunity to study in his institute, first as a master student and then for my PhD. I'm greatly appreciative of the support, guidance, and motivation throughout my studies. I would also like to thank Prof. Dr. Julian Caspers for his guidance and assistance as my secondary supervisor during my PhD project.
A particular thanks goes to PD Dr. Susanne Weis for assisting me in all aspects of my PhD. Thank you for your patience and for encouragement in times of need. Your supervision has helped me to become an independent researcher and prepared me for my future career in science. An additional thanks goes to all my colleagues from INM-7 who have always been very kind and made my PhD more enjoyable. I want to particularly thank the people from my working group Dr. Lisa Mochalski, Dr. Patrick Friedrich, Dr. Xuan Li, Dr. Julia Camilleri, Dr. Lisa Wiersch, Gianna Kuhles and Natalie Schlothauer for supporting me throughout my PhD and my Düsseldorf office colleagues, Kaustubh, Jürgen, Leo, Tobias, Georgios, Mamaka and Sam for the interesting lunch time discussions. I also want to thank Anna and Julia. Thank you both for bringing fun and joy to my years at the INM-7 and thanks for putting up with me.

A special thanks goes to:
Nevena, from bachelor's students to PhDs - what a ride! Thank you for making the PhD a fun (and almost enjoyable) experience. Although you are never on time, you are always there when I need you. Tobi, you never answer my texts - but I know you love me. Thank you for always supporting me, no matter what. Felix, you have enough on your plate as a BVB fan so I will **not** say "Ha, I made it before you!". Thank you for always listening to my problems, and sometimes even share personal information yourself, you're a great friend.

To my family. Thank you, Mama & Papa for all the trust and confidence you have in me. Thank you for all the support throughout my studies. I could not have done this without you. Thank you, Freddy, for making life fun even in the most challenging times. I know I can always count on you.

Most importantly, I want to thank my wife, Janina. Thank you for being my partner in crime. Your unwavering support has helped me through this PhD (and through life in general if I'm being honest). You and Merle make me the happiest man alive. I will forever be grateful for everything you have done for me.

Disclaimer: As of now, my heartfelt thanks are extended to the children I currently have. If, at any future point, additional children should exist, he/she is to be considered equally acknowledged and appreciated, and any expressions of heartfelt thanks should be understood to include him/her as well. Likewise, any friends who may not have been specifically mentioned are nonetheless sincerely valued and included in my gratitude.