DTU Library

Neuroimaging based predictions and subtyping in Schizophrenia Spectrum Disorders

Krohne, Lærke Gebser

Link back to DTU Orbit

Lærke Gebser Krohne

NEUROIMAGING BASED PREDICTIONS AND SUBTYPING IN SCHIZOPHRENIA SPECTRUM DISORDERS

PhD Thesis, March 2023

DTU Compute
Department of Applied Mathematics and
Computer Science

Neuroimaging based predictions and subtyping in Schizophrenia Spectrum Disorders

PhD dissertation by:
Lærke Gebser Krohne

Main Supervisor:
Kristoffer H. Madsen, Associate Professor, DTU Applied Mathematics and Computer Science

Co-supervisor:
Søren R. Christensen, Lundbeck A/S, Department of Clinical Pharmacology

DTU Applied Mathematics and Computer Science
Section of Cognitive System
Technical University of Denmark
Richard Petersens Plads 321, 2nd floor
2800 Kgs. Lyngby
Denmark
Tel: +45 45 25 30 40

ABSTRACT

Schizophrenia is a complex neuropsychiatric syndrome with a high internal heterogeneity (inter-individual variations in neurobiological, genetic and phenotypic profile). Currently, the causes and neurobiology of schizophrenia are not fully understood, and there is a large unmet medical need, since many patients do not respond adequately to available treatments and have poor long term outcomes. Even though various mechanistically plausible biomarkers for schizophrenia have been suggested, none of these are so far used clinically. Having reliable and objective biomarkers is important to better understand the disorder, to support clinical decision and to assist the development of new treatments.

Early studies show that patients with schizophrenia have widespread impairments in neural communication, which can be measured using functional magnetic resonance imaging (fMRI). Abnormal brain activation has been linked to different aspects of the disorder, but firm conclusions are yet to be made, since the results vary across studies. These variations are often attributed to the high internal heterogeneity and limited sample sizes in most studies. In the last decade, the use of data-driven methods and large multi-site datasets, have led to several advancements, which is promising for future research progress. However, overall the field is still at a stage of identifying solutions to methodological challenges, rather than developing specific biomarkers for clinical practice (yet).

The goal of this PhD project was to address a part of the methodological puzzle, by exploring different ways to use machine learning in the search for robust and reproducible fMRI biomarkers. Throughout the analyses, we have used supervised machine learning to enable clinical predictions and unsupervised machine learning for feature extraction (decomposition methods) and disease subtyping. The work has been organized into four studies as summarized below.

In an attempt to search for early risk prediction biomarkers, the goal of Study 1 was to classify healthy participants with schizotypal traits according to their degree of social anhedonia. We developed a classification framework with a broad selection of feature extraction methods to determine which of these could drive the classification. We found significant predictions when using both temporal and spatial network features, and achieved the highest performances, when using features from the two data-driven decomposition methods: independent component analysis (ICA) and multi-subject archetypal analysis (MSAA). Throughout our analyses, we discovered how much the final results depended on the parameters within the analysis pipeline. Thus, for the remaining studies, we focused our analysis to increase the robustness and reproducibility, e.g. by using multi-site data which had been made publicly available through data-sharing initiatives. This enabled us to train the models on a more heterogenous multi-site discovery dataset, and to test the generalizability of our findings on external data.

In Study 2 the goal was to classify patients with schizophrenia using multi-site data. We adjusted the prediction framework to handle multi-site predictions and furthermore aimed to make each step as data-driven and robust as possible. For the decomposition methods, we also

investigated different ways of using transfer learning to bridge feature extraction between datasets. Using spatial network features from both the decomposition methods and parcellation-based connectivity analysis we found high and reproducible classification performances that generalized to the external data. The highest performances were obtained when using ensemble decision models, which supports earlier findings that schizophrenia affects a wide range of brain networks.

In Study 3, we used the same features to predict the symptom severity (measured using the Positive and Negative Syndrome Scale (PANSS)) and three PANSS subscales in an attempt to address the internal heterogeneity of schizophrenia. We used Gaussian process regression (GPR), which is a non-parametric Bayesian approach to regression that provides an uncertainty estimate for the predictions. Here, we only found moderate prediction performances, which resembled a positive trend around the mean PANSS score, and which generally did not reproduce on the external data. These findings indicate that the study could be underpowered or that the between-site differences are too large compared to the signal of interest. Another possible explanation could be the internal consistency of the PANSS scales, or that the used datatype (resting state connectivity) or applied methods are not the right path forward.

Finally, in Study 4, the goal was to search for data-driven disease subtypes using a multiple co-clustering (MCC) method that is based on Bayesian mixture models. Since the subtyping field is still at an exploratory stage, we dedicated a large part of our investigation to study the stability of the MCC method. We found that the clustering solutions were highly dependent on changes in the dataset, Nevertheless, we found subtypes with significant diagnosis association that reproduced on the external data.

To conclude, we see our work providing important methodological contributions towards using machine learning and multi-site data in the search for robust and reproducible fMRI biomarkers. All our analyses were performed on fMRI data either from individuals at risk of developing psychosis (Study 1) or from patients with schizophrenia (Studies 2 − 4), however the developed methods can be directly used to study other clinical populations.

RESUMÉ

Skizofreni er en kompleks neuropsykiatrisk lidelse, og mange patienter har begrænset behandlingsrespons og dårlig langsigtet prognose. I dag bliver patienter med skizofreni primært diagnosticerede og evaluerede i forhold til graden af deres kliniske symptomer. En stor udfordring i forhold i denne forbindelse er den høje interne variation i patienternes sygdomsbillede, som gør at patienter, til trods for samme diagnose, er påvirkede af meget forskellige symptomer, mens der samtidigt er et betydeligt overlap mellem symptomer på tværs af andre psykiatriske lidelser. I de seneste år er der derfor blevet søgt efter nye sygdomsdefinitioner samt 'sygdomsundertyper', som muliggør at patienter kan indeles i grupper med mere homogen biologi. Her spiller biomarkører, som er definerede som objektive biologiske målinger, en vigtig rolle, da de har potentialet til at overkomme de tidligere udfordringer.

Tidligere studier har vist, at patienter med skizofreni har udbredte svækkelser i den måde som forskellige områder at hjernen kommunikerer med hinanden på, hvilket kan måles ved hjælp af hjerneskanningsmetoden funktionel magnetisk resonans-billeddannelse (fMRI). Selvom atypisk hjerneaktivitet er blevet forbundet med mange forskellige aspekter af lidelsen, er der stadig meget variation på tværs af studier i forhold til resultater og metoder. Dette er ofte tilskrevet til den høje interne heterogenitet af lidelsen (f.eks. inter-individuelle variationer i symptomer, neurobiologi og genetik) samt at datasættene for de enkelte analyser har været af begrænenset størrelse. I de seneste år har brugen af datadrevne metoder samt store datasæt, der er opsamlet flere forskellige steder (multi-site), ført til hurtige fremskridt. Overordnet er feltet dog stadig i gang med at finde løsninger på metodiske udfordringer frem for at bidrage med specifikke biomarkører til klinisk brug.

Formålet med dette ph.d.-projekt var at udforske forskellige måder at bruge machine learning (maskinlæring) i søgen efter robuste og reproducerbare fMRI-biomarkører. Vi har brugt supervised machine learning til at kunne prædiktere kliniske mål (som diagnose og symptomsværhedsgrad) og "unsupervised machine learning" til at 'udtrække' hjernenetværk og til at finde nye sygdomsundertyper. Overordnet set har vi opdelt vores arbejde i fire studier.

Formålet med Studie 1 var at klassificere raske deltagere med skizotypiske træk i henhold til deres grad af social anhedoni, og herved søge efter biomarkører, der kan estimere risikoen for at udvikle skizofreni. Her undersøgte vi en række forskellige måder til at udtrække informationer fra hjerneområder og -netværk, blandt andet ved at bruge de datadrevne dekompositionsmetoder: independent component analysis (ICA) og multi-subject archetypal analysis (MSAA). Studiet viste at det var muligt at opnå signifikant prædiktioner, når vi brugte temporale og rummelige netværksinformationer, mens dette ikke var muligt med stationære hjerneaktivitetsmål. Yderligere observerede vi at de endelige resultater afhang meget af paramterne for de enkelte analyser. For de resterende studier fokuserede vi derfor på at øge robustheden og reproducerbarheden, f.eks. ved at bruge multi-site datasæt, som var blevet gjort offentligt tilgængelige. Dette gjorde det muligt for os at udvikle vores modeller på mere heterogene multi-site datasæt og at teste generaliserbarheden

af vores resultater på nye datasæt.

I Studie 2 var formålet at klassificere patienter med skizofreni på et multi-site datasæt. Vi justerede vores klassifikations-modeller til at håndtere multi-site data og sigtede desuden efter at gøre hvert trin så datadrevet og robust som muligt. For dekompositionsmetoderne undersøgte vi også forskellige måder at bruge transfer learning til at bygge bro mellem datasæt. Vi fandt høje og signifikante klassifikationsresultater, som også var signifikante, når vi testede dem på det nye datasæt. De bedste resultater blev opnået ved brug af ensemblebeslutningsmodeller, som inkluderede information fra alle hjernenetværk, hvilket understøtter tidligere studier, som har vist at skizofreni påvirker mange forskelle hjernenetværk.

I Studie 3 brugte vi de samme hjernenetværksinformationer til at prædiktere symptomernes sværhedsgrad (målt ved hjælp af Positive and Negative Syndrome Scale (PANSS)) og de tre PANSS subskalaer i et forsøg på at adressere den interne heterogenitet af skizofreni. Vi brugte Gaussisk process regression (GPR), som er en ikke-parametrisk Bayesiansk tilgang til regression, der giver et usikkerhedsestimat for prædiktionerne. Her fandt vi dog kun moderat nøjagtighed af prædiktionerne, som repræsenterede en positiv korreleret tendens omkring den gennemsnitlige PANSS-værdi, og som overordnet set ikke generaliserede sig til det nye datasæt.

I Studie 4 var formålet at søge efter datadrevne sygdomsundertyper ved hjælp af metoden Multiple co-clustering (MCC), som er en segmenteringsmetode (clustering), der er baseret på Bayesiaske mixture models. Da feltet stadig er nyt og i udvikling, dedikerede vi en stor del af vores analyser til at undersøge stabiliteten af MCC-metoden. Vi fandt ud af, at resultatet var meget afhængigt af ændringer i datasættet. Ikke desto mindre fandt vi en subtype-løsning, som havde en signifikant diagnosesammenhæng der også blev genfundet på de eksterne data.

Vi mener at vores arbejde bidrager med vigtige metoder til at bruge machine learning og multi-site data i søgen efter robuste og reproducerbare fMRI-biomarkører. Alle vores analyser blev udført på fMRI data fra enten personer med risiko for at udvikle skizofreni (Studie 1) eller patienter (Studie 2-4). Det er dog også muligt at anvende de udviklede metoder direkte til at undersøge andre kliniske populationer.

PREFACE

The work of this PhD project comprises four studies, which have been carried out during two enrollment periods. I initially started my PhD in July 2017, where I carried out Study 1 (resulting in Paper A and B) before going on maternity leave in the beginning of 2019. After my maternity leave, I started a new job at H. Lundbeck A/S (2-year Graduate program based in the Clinical Biomarker department). From September 2020 to December 2021, I was therefore not enrolled as a PhD student, but worked full time at H. Lundbeck A/S. In December 2021, I then re-started my PhD, this time in a collaboration between DTU and H. Lundbeck A/S, where I carried out Studies 2-4 (resulting in Paper C and D). The total PhD enrollment period is therefore 3 years and 3 months, even though the project was started in 2017.

This PhD thesis consists of a brief introduction to the overall topics within the field, description of the applied methods, summary of the research contributions and a discussion of these. The thesis furthermore consists of four papers, of which two are accepted and two are in preparation.

The PhD project has been carried out at the section of Cognitive systems at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark, Kongens Lyngby, Denmark. During my first enrollment I was also affiliated at the Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Denmark. During my second enrollment I have been fully employed at the Clinical Biomarker group at H. Lundbeck A/S. Five months (April 2022–August 2022) were used for an external research stay in the Data science group at the Department of Biometrics at H. Lundbeck A/S.

The research was collaboratively funded by H. Lundbeck A/S and the Department of Applied Mathematics and Computer Science at the Technical University of Denmark and.

Lærke Gebser Krohne

CONTRIBUTIONS

Journal papers

- Madsen, Kristoffer H; Krohne, Laerke G ; Cai, Xin-Lu; Wang, Yi; Chan, Raymond CK; **Perspectives on machine learning for classification of Schizotypy using fMRI data**. *Published in Schizophrenia Bulletin, 2018.*

- Krohne, Laerke G ;Wang, Yi; Hinrich, Jesper L; Mørup, Morten; Chan, Raymond CK; Madsen, Kristoffer H; **Classification of social anhedonia using temporal and spatial network features from a social cognition fMRI task**. *Published in Human Brain Mapping, 2019.*

- Krohne, Laerke G ;Christensen, Søren R; Mørup, Morten; Madsen, Kristoffer H; **Neuroimaging based predictions of Schizophrenia diagnosis and PANSS scores, a multisite resting state fMRI study** . *In preparation*

- Krohne, Laerke G ;Hansen, Ingeborg H; Madsen, Kristoffer H; **On the search for data-driven and reproducible schizophrenia subtypes using resting state fMRI data from multiple sites**. *In preparation*

Conference abstracts

- Krohne, Laerke G ;Christensen, Søren R; Mørup, Morten; Madsen, Kristoffer H. **Multi-Subject Archetypal Analysis on multi-site rsfMRI data for classification of Schizophrenia**. *Presented at the Organization for Human Brain Mapping (OHBM) annual meeting 2022, Glasgow, Scotland.* Peer reviewed abstract and poster presentation (both virtual and in person)

Peer reviewed papers not included in thesis:

- Karabanov, Anke N ;Madsen, Kristoffer H; <u>Krohne, Laerke G</u>; Siebner, Hartwig S; **Does pericentral mu-rhythm "power" corticomotor excitability? - A matter of EEG perspective**. *Published in Brain Stimulation , 2021.*

- Madsen, Kristoffer H; , Karabanov, Anke N ; <u>Krohne, Laerke G</u>; Safeldt, Mads G; Tomasevic, Leo; Siebner, Hartwig S; **No trace of phase: Corticomotor excitability is not tuned by phase of pericentral mu-rhythm**. *Published in Brain Stimulation , 2019.*

- Hansen, Sofie T; Hemakom, Apit; Safeldt, Mads G; <u>Krohne, Laerke G</u>; Madsen, Kristoffer H; Siebner, Hartwig S; Mandic, Danilo P; Hansen, Lars K; **Unmixing oscillatory brain activity by EEG source localization and empirical mode decomposition**. *Published in Computational Intelligence and Neuroscience , 2019.*

ACKNOWLEDGMENTS

To all my family and friends: thank you for always listening and supporting me, both when I wanted to explain the "super exciting thing I just discovered" and when the project was difficult and things did not went the way that I had hoped.

And last but not least, my greatest thanks and gratitude to my husband *Marcel* and daughter *Freya*. Thank you for all your support, both when I had to work late, again and again, but also for helping me to set boundaries and put aside work (for a while). You have taught me that high quality and temporal duration are not only important to achieve reliable MRI readouts, but also core aspects of taking breaks. Thank you for all the joy and love you both bring into our lives, even when times are hard. I love you to the moon and back.

TABLE OF CONTENTS

LIST OF FIGURES

NOMENCLATURE

AUC   area under the curve

BOLD   blood-oxygen-level-dependent

CD   Cohen's D

CV   cross validation

dATT   dorsal attention

DMN   default mode network

DSM   Diagnostic and Statistical Manual of Mental Disorders

EPI   echo-planar imaging

fMRI   functional magnetic resonance imaging

FPN   fronto-parietal network

GICA   group indpendent component analysis

GP   Gaussian process

GPR   Gaussian process regression

HC   healthy control

ICA   independent component analysis

ICD   International Classification of Diseases

MDL   minimum description length

ML   machine learning

MNI   Montreal Neurologic Institute

MRI   magnetic resonance imaging

MSAA   multi-subject archetypal analysis

PANSS   Positive and Negative Syndrome Scale

PCA   principal component analysis

PDF   probability density function

PET   positron emission tomography

RDoC   Research Domain Criteria

RFT   random field theory

ROC   receiver operating characteristic

ROI   region of interest

RSN   resting state network

SNR   signal-to-noise ratio

SoMo   somoatomotor network

SPM   Statistical Parametric Mapping

SVM   support vector machine

SZ   schizophrenia

vATT   ventral attention

INTRODUCTION

The brain is often considered to be the last frontier in biological sciences, and even though the field of neuroscience has made enormous progress over the past decades, much is still unknown. This has large implications for our opportunity to understand and develop treatments for patients suffering from brain disorders, which impact more than 3 billion people worldwide [1, 2]. Traditionally psychiatric disorders are diagnosed based on their symptomatology using diagnostic tools such as the Diagnostic Manual of Mental Disorders (DSM) [3]. Additionally, clinical scales are used to measure disease related phenotypic measures, such as symptom severity, disease progression and treatment responses. These are subjective measures, which are either determined by the patient or a health care professional. A core challenge for these measurements is the high internal heterogeneity of psychiatric disorders, implying that patients with the same diagnosis can be affected by symptoms in several different domains, while there are also a substantial overlap between symptoms across disorders. Several initiatives, such as the Research Domain Criteria Project (RDoC) [4, 5], have therefore been established to find more mechanistic disease definitions or even disease subtypes with more homogeneous biology. Here biomarkers, which are objective measurements of biological processes, play an important role, as they have the potential to overcome earlier challenges of symptomatology based measures [6, 7].

Schizophrenia is a complex neuropsychiatric syndrome with a high internal heterogeneity, both within the neurobiology, genetics and symptomatic profile. The symptom severity is often assessed using the Positive and Negative Syndrome Scale (PANSS), which is further organized into three subscales: positive, negative and general psychopathology, reflecting some of the core symptom domains of the disorder [8]. Whereas positive symptoms (such as hallucinations) are generally effectively managed with antipsychotic medications [9], treatment options for negative symptoms, such as anhedonia (the reduced ability to experience pleasure), and cognitive deficits are limited [10, 11]. So far, there are no clinically used biomarkers [6], but they would be a valuable tool for many applications, such as supporting diagnostic decisions, and in drug development (where it has been suggested that the high internal heterogeneity and the use of clinical scales to measure efficacy in clinical trials, have haltered the development of new treatments [12]). Neuroimaging can be used to measure important biological processes, ranging from activation of brain receptors to network interactions between brain regions, and is thus a strong candidate for biomarker discovery. Functional magnetic resonance imaging (fMRI) is an indirect measure of brain activation, which is often used to study activation in specific brain regions or in brain networks. Many studies have already used fMRI to search for neurobiological underpinnings that can support a range of clinical applications, but firm conclusions are still to be made, since earlier studies have shown substantial differences both in their applied methods and results [13–15]. In recent years, the field has started moving towards using more data-driven methods and larger multi-site datasets [16, 17]. This carries a great potential to overcome earlier challenges and advance fMRI biomarkers into clinical applications and drug development [15–18].

1.1   Structure and objectives

The overall aim of this PhD project was to explore ways of using machine learning to search for data-driven and reproducible fMRI biomarkers for patients with schizophrenia. We have investigated supervised machine learning methods to make clinical predictions, and unsupervised machine learning for feature extraction and disease subtyping. Throughout our analyses, we have focused on robustness and reproducibility.

The objectives of this PhD project were to:

- Explore ways of using feature extraction and supervised machine learning to obtain reproducible predictions of phenotypic measures, such as diagnostic labels and outcomes from clinical scales.
- Compare different feature extraction methods, hereby data-driven decomposition, with regards to their stability, interpretability and predictive performance
- Investigate how multi-site data can be used to search for robust biomarkers that generalize across datasets
- Search for data-driven subtypes with a more homogeneous biology by using unsupervised clustering and investigate the stability of the clustering solutions

Studies included in this PhD project

The PhD work was structured into four studies as illustrated in Figure 1.1.
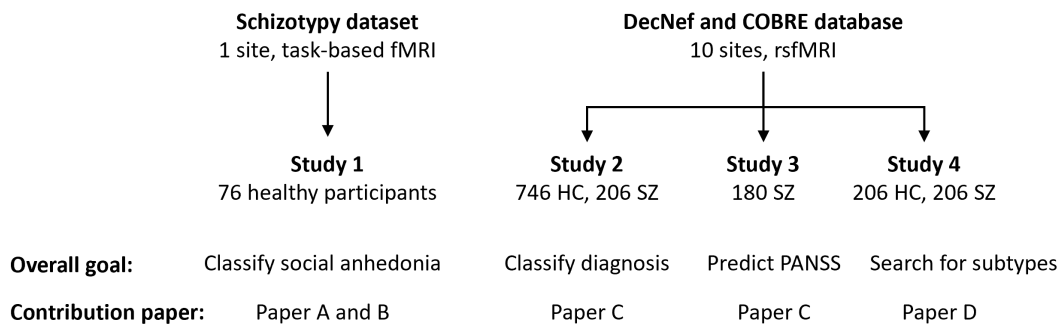
Figure 1.1: Overview of four studies. Brief overview of the main goals, datasets and corresponding papers for the four studies of this PhD project. The DecNef and COBRE databases and the way the data differed between the studies are described in section 2.6.

Study 1 was performed on a single-site dataset that included healthy participants with varying degree of schizotypy (set personality traits that are related to schizophrenia). For the remaining studies, we used multi-site fMRI data from patients with schizophrenia (SZ) and healthy controls (HC), which were made publicly available between my two PhD enrollments (as described in the preface). We used data from the Decoded Neurofeedback (DecNef) Project Brain Data Repository [19] and the Center of Biomedical Research Excellence (COBRE) [20] databases.

The main goal(s) of each study was to:

- **Study 1**: Classify participants with high and low anhedonia, by building a classification-framework with various feature extraction methods including the novel multi-subject archetypal analysis (MSAA) decomposition method
- **Study 2**: Classify participants according to their diagnosis (SZ or HC) using multi-site data, and extend the classification-framework to focus more on robustness and reproducibility
- **Study 3**: Predict the symptom severity (using the total PANSS score) and three PANSS subscales in an attempt to address the internal heterogeneity of schizophrenia
- **Study 4**: Search for data-driven subtypes with a more homogeneous biology by using clustering, and evaluate the stability of the clustering solutions

More details on the goals and datasets used in all four studies can be found in section 2.5.

The four studies had a similar structure for biomarker discovery as illustrated in Figure 1.2. First the raw data was converted into interpretable brain features (using prepossessing and feature extraction), which were subsequently used for predictive modelling and subtyping to search for disease related biomarkers.
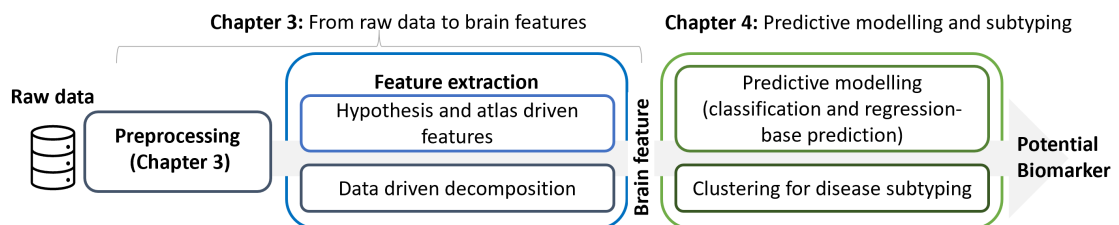
Figure 1.2: Overall framework for all four studies. Throughout all four studies we used preprocessing and feature extraction to obtain brain features, which were used for subsequent prediction or subtyping.

1.1.1 Thesis outline

This thesis consists of four papers (two accepted and two in preparation), and several chapters that tie together the contributions. Chapter 2 gives a brief introduction to the overall topics of this thesis: schizophrenia, fMRI imaging for biomarker discovery, machine learning and multi-site imaging. Chapter 3: "From raw data to brain features" describes how preprocessing and feature extraction was used throughout the studies, and Chapter 4 describes how we used machine learning for predictive modelling and disease subtyping. In Chapter 5 the research contributions are summarized, and the overall contributions and future perspectives are discussed in Chapter 6.

BACKGROUND

2.1 Brief introduction to Schizophrenia

Schizophrenia is a psychiatric syndrome with a complex and heterogenous neurobiological, genetic, and phenotypic profile [6, 21], which affects approximately 24 million (0.32 %) people world wide [22, 23]. No single cause of schizophrenia has been identified, but it is believed to be caused by a complex interplay of genetic and environmental risk factors, which influence the brain development and the biological adaptation to life's experiences [24]. These risk factors include: prenatal events (e.g. complications during pregnancy or delivery), gender (more frequent in men (1.4/1), which also have an earlier onset and more severe symptoms [24, 25]), drug abuse and social adversity [24, 26]. This is not an exclusive list, and schizophrenia is also highly heritable [27] where genetics constitute an important and complex risk factor [27–29]. The diagnosis of schizophrenia is based on the symptomatology using diagnostic tools, such as the the Diagnostic Manual of Mental Disorders (DSM) [3] or the International Classification of Diseases (ICD)[30], which evaluates the severity and duration of the symptoms and other aspects such as social and occupational functioning, comorbidity, and substance abuse [24]. Characteristic symptoms of schizophrenia are commonly divided into positive, negative and cognitive categories. Positive symptoms are behaviours and thoughts which are not normally present in healthy individuals, such as hallucinations and delusions. Negative symptoms include social withdrawal, diminished initiative and energy as well as anhedonia (inability to feel pleasure). Lastly, cognitive disturbances span a broad set of cognitive dysfunctions such as problems with attention, memory and reasoning.

The onset of schizophrenia is often diagnosed during late adolescence and most patients experience serious impairment in many domains of everyday life [31, 32]. Treatments mainly include antipsychotic medications (all currently licensed treatments for schizophrenia are currently D_2 blockers[29]) and psychosocial interventions,such as cognitive-behavioural therapy [24]. Unfortunately there is still a large unmet medical need, since many patients do not respond adequately to currently available treatments (particularly for the negative and cognitive dimension) and have severe side effects[33, 34]. Furthermore, even though treatments improved the quality of life, and enables most patients to live on their own, the effect on social and professional functioning is still limited [24, 29].

Two of the major hypotheses for the pathoetiology of schziophrenia relate to the importance of i) the dopaminergic dysregulation [35] and ii) viewing schizophrenia as a neurodevelopmental disorder where the pathogenesis begins already during early development[1] [26, 37]. The former highlights the importance of the dopaminergic neurotransmitter system where earlier studies using positron emission tomography (PET) and amphetamine- fMRI studies have shown dopamine

---

[1]Recent studies have argued that the two hypothesis may be integrated with previous work on the excitation-inhibition balance of schizophrenia [36]

dysregulation, particularly in the striatum [38, 39]. Furthermore, other studies have shown that schizophrenia is associated with circuit-level alterations including both cortical and subcortical regions[6, 40–42], which can be measured with fMRI as described in section 2.5.2. The improved neurodevelopmental understanding has further highlighted the importance of finding early onset symptoms and biomarkers. One way to study the early course of the disorder is to look at individuals with schizotypal traits. Schizotypy refers to a set personality traits (organized into positive, negative and generalized) that are related to schizophrenia [43]. Individuals with schizotypy have psychotic-like experiences, which have shown to reflect their vulnerability for developing schizophrenia later in their life [44–46] as described in section 2.5.3.

2.1.1   The Positive and Negative Syndrome Scale (PANSS)

The Positive and Negative Syndrome Scale (PANSS) is a clinical scale that is often used to measure the symptom severity in patients with schizophrenia. The scale consists of 30 items, which are grouped into a positive (7 items), negative (7 items) and general psychopathology subscale (16 items, this subscale is also referred to as "generalized" in the remaining thesis)[47]. Patients are rated from $1 - 7$ (ordinal scale) for each of the 30 items, based on a clinical interview as well as reports from family members or primary health care professionals [48]. The PANSS scale is widely used in clinical practice to measure outcomes such as disease progression and treatment response (also in clinical trials). Furthermore, the PANSS scale has been used to identify psychopathological subtypes, in an attempt to disentangle the heterogeneity in schziophrenia, however so far with inconclusive results [49, 50].

2.1.2   Biomarkers

A biomarker is defined as an biological indicator of a biological process (normal or pathological) or response to an exposure or intervention [7]. Typically, the purpose of a biomarker is to give an objective measure that can support clinical decisions. Figure 2.1 illustrates some of the relevant types of biomarkers for schizophrenia. The line indicates the degree of symptoms over time, ranging from no symptoms, to a prodromal phase (increasing degree of symptoms but below diagnostic threshold) until the person is diagnosed with schizophrenia (active phase). Examples of biomarker types include: 'risk prediction biomarkers' to identify individuals at risk for psychosis, 'diagnostic biomarkers' to assist diagnostic decisions, and 'therapeutic monitoring biomarkers', to monitor the effectiveness of a treatment. During drug development, 'pharmacodynamic biomarkers' can be used to evaluate the biological effect of the pharmacological intervention (elaborated in section 2.5.1) .

Currently there are no clinically used biomarkers that can inform diagnostic and treatment decisions in schizophrenia (SZ) [6, 15], but neuroimaging is a strong candidate for biomarker discovery as recently described by Kraguljac et al. [6]. The list of potential neuroimaging biomarkers covers a broad range, such as altered release of neurotransmitters, receptor occupancy, neuroinflammation and dysconnectivity between brain regions [6]. In this PhD project we focused on data from functional magnetic resonance imaging for biomarker discovery.
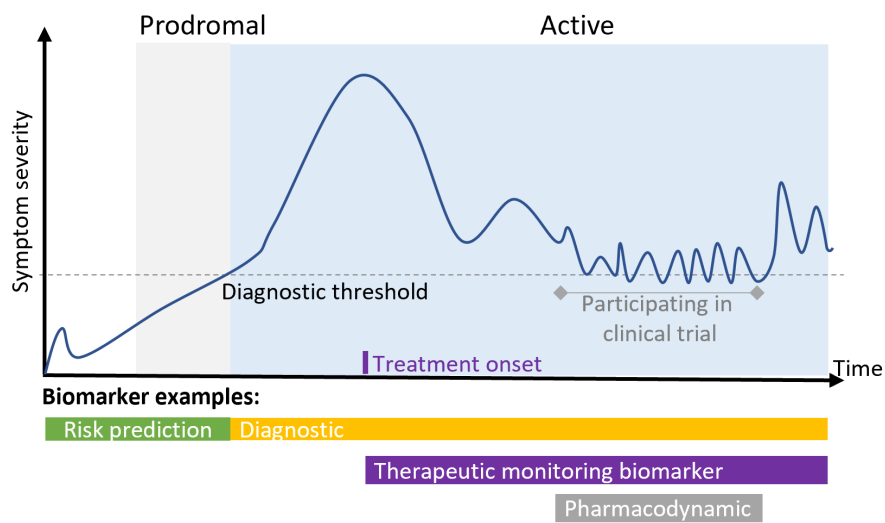
Figure 2.1: Examples of biomarker types in schizophrenia. This figure illustrates the symptom severity over time and examples of when different types of biomarkers would be relevant . The line indicates the symptom severity over time. Commonly schizophrenia is described to have a prodromal phase, where symptoms increase but are below a diagnostic threshold, and an active phase which is the time after the symptoms have crossed the threshold. The figure lists examples of relevant biomarkers, however this is not a comprehensive list. This figure was inspired by Figure 2 from Kaguljac et al. in 2021 [6]

2.2 Functional MRI and brain connectivity

Functional magnetic resonance imaging (fMRI) is a non-invasive brain imaging technique that can be used to study brain activation either during a specific task or during rest. Mostly fMRI imaging leverages the blood oxygen level dependent (BOLD) contrast [51]. In short, when the activation level of a brain regions changes, this will influence its energy consumption and thereby change the the presence of oxygenated blood. Due to the magnetic properties of hemoglobin in the blood, this can be measured with an MRI scanner [51, 52]. The BOLD signal is thus an indirect measure of brain activation, which is related through the haemodynamic response function. Whereas other fMRI method exist (e.g. arterial spin labelling [53]), these will not be described in this PhD Thesis. Compared to other functional brain imaging methods (such as electroencephalography), fMRI is limited to a lower temporal resolution due to the temporal smoothness of the haemodynamic response function (in the order of seconds). However, in many applications it is still preferred to other non-invasive functional imaging techniques since it has higher spatial resolution (1-4 mm^2) and subcortical brain areas are accessible for imaging.

When fMRI is used to study brain activation during a certain **task**, a control condition is needed to contrast the brain activation between these two conditions. Traditionally, task based brain activation studies have relied of univariate brain mapping approaches, where brain activation in isolated brain regions are related to an phenotypic measure (observable characteristics of an individual) of interest. Examples of health-related phenotypic measures include diagnostic labels (e.g. healthy controls vs patients with SZ), symptoms severity, and cognitive scores. In this way,

univariate mapping builds on a historical foundation of lesion studies [54], where the goal was to understand the functions and processes encoded in isolated brain regions [17].

In 1995, Biswal and colleagues discovered that distinct brain regions exhibit synchronous fluctuations in intrinsic activity which give rise to so-called **functional connectivity (FC)** between regions. Given the FC patterns, several temporally coherent networks have been found which sub-serve critical functions such as audition, vision, motor planning and directed attention [55–57]. These networks have been shown to be surprisingly consistent (though not identical) patterns of activation both during tasks and resting state [57, 58], and are often acquired while the participant is resting with their eyes open or closed while not performing any explicit task in the scanner. The large scale intrinsic brain networks are commonly referred to as resting state networks (RSN), and they have been associated to many important functions and diseases[59, 60]. One RSN parcellation which is commonly used in the literature is the 7-RSN parcellation that was presented by Yeo et al. in 2011, as illustrated in Figure 2.2.
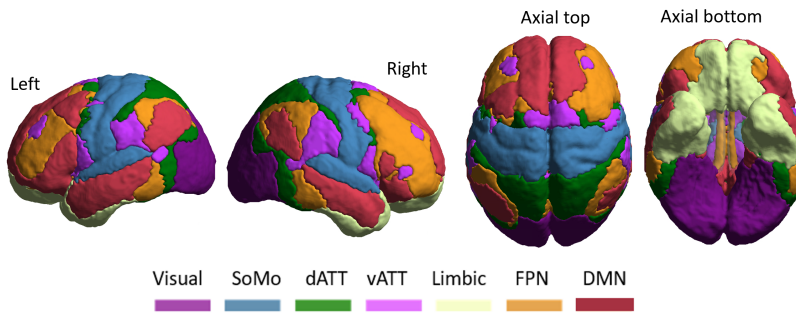
Figure 2.2: Resting state network (RSN) parcellation. Illustration of the 7-RSN parcellation presented by Yeo et al. in 2011. This RSN parcellation is widely used in the literature as well as in Studies 2-4 of this PhD project. The seven RSN include: Visual, Somatomotor (SoMo), dorsal attention (dATT), ventral attention (vATT), Limbic, Frontal-parietal (FPN) and Default mode network (DMN).

So far, univariate brain mapping and functional connectivity studies have enabled much progress for 'traditional neuroscience questions', e.g. fMRI has made large contributions to the field of understanding cognition, such as fundamentally changing how we think about about the aging mind [61–63]. However, there are only few fMRI applications (mainly in presurgical mapping) in clinical practice so far [17, 64].

As for many fields in medical science and psychology, fMRI has been plagued by challenges of reproducibility across studies [64, 65]. Main culprits for the reproducibility challenge have been attributed to the intrinsic reliability of fMRI signal, the large number of parameters that differ between studies (in acquisition, preprocessing, statistical analysis and reporting) and relatively limited sample sizes [64, 66]. These topics will be a main focus throughout the thesis.

Applications of fMRI have been particularly challenging in psychiatric brain disorders, where the heterogeneity of the disorders and the lack of precision in diagnostic categorizations have further challenged the reproducibility between studies [15].

For the last decade, the focus has therefore been on the development of new analytical and more data driven techniques (section 2.3) and large multi-site datasets (section 2.4), with the hope

that these can overcome earlier limitations and transform the use of neuroimaging for clinical applications [15, 17].

2.3 Machine learning in neuroimaging

The overall goal of machine learning is to make machines classify data without being explicitly programmed. Typically machine learning methods are categorized into supervised and unsupervised methods. Supervised methods aim to build a model that can predict a measure of interest (label), in contrast unsupervised methods explore statistical dependencies in unlabeled data to learn structures. Here, we will give a brief introduction to the overall concepts, while more detailed descriptions of the machine learning methods that were used across our studies are given in chapter 4.

Figure 2.3 is an adjusted version of Figure 1.2, where we have specified in which parts of the PhD project we have used machine learning methods (red).
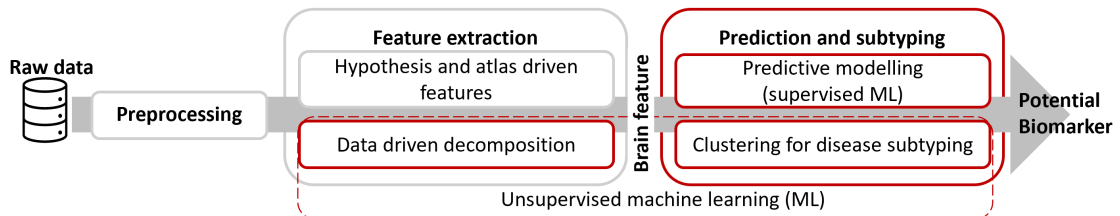
Figure 2.3: Use of machine learning throughout the studies. This figure is an adapted version of Figure 1.2, which highlights parts of our data analysis where we used machine learning (marked with red). The remaining (gray) steps were performed using more traditional fMRI analytical approaches.

2.3.1 Supervised machine learning methods

These methods are conceptually similar to conventional brain mapping analysis (described in section 3.2) but there are some important differences as specified by Woo et al. [17]:

1. the direction of inference is reversed such that the brain features are a set of predictors while the label comprise one or more outcomes
2. the models include all available data features (multivariate) to make a single prediction about the outcome
3. the models can capture information on multiple spatial scales and datatypes (e.g. combining functional and structural imaging)
4. the diagnostic value of predictive models are evaluated on 'out-of-sample' individuals, which means that the data that was used to train the model did not include the individuals on which its performance was tested. This can be done either using cross-validation, or by leaving out a part of the data for independent testing (preferably on external data which comes from independent datasets) if sufficient data is available.

Whereas most fMRI studies so far have focused on **binary classification** (e.g., between diagnosis labels such as SZ and healthy controls [6, 17]), an emerging trend is to predict individual differences

in continuous outcomes, such as symptom severity or cognitive functions [16]. Compared to binary classifications, these **regression-based predictions** can be more challenging, but have the potential to shed light on more fine-grained differences, e.g. on the trajectory of brain alternations with fluctuating symptom severity. Furthermore, they can be used to reduce the internal heterogeneity which is inherent in most psychiatric diagnoses [16, 21].

A challenge of predictive modelling is that the model training and evaluation depends on the phenotypic 'labeling', which is considered the gold standard. In psychiatry this can be challenging, since even though considerable effort are put into standardizing clinical measurements, the reliability of the labelling may not always be high [67, 68]. This can furthermore be a challenge for multi-site dataset where different raters (and potentially even rating systems) have been used.

2.3.2 Unsupervised machine learning

The overall goal of unsupervised machine learning is to discover statistical dependencies in the data without using labeled information. For fMRI data, two commonly used applications are i) decomposition methods for feature extraction and ii) clustering for disease subtyping.

Feature extraction: Due to the high dimensional of fMRI data, feature extraction is an important step that is included in most fMRI studies. Often, feature extraction is approached by either using brain atlases (parcellation based approaches) or by unsupervised decomposition methods, which can find data-driven patterns in the data. The most commonly used decomposition method for feature extraction in fMRI data is independent component analysis (ICA), which aims to identify a latent representation of the data, such that the sources are maximally independent. In practice, methods such as ICA find brain regions with consistent temporal fluctuations, which means that they can be used to extract functional connectivity networks.

Subtyping: Unsupervised clustering methods can be used for disease subtyping, i.e. to find subtypes within the fMRI data with a more homogeneous biology. Even though fMRI based disease subtyping has been a goal for many years, the subtyping field has previously been challenged by the relatively small sample sizes and the high dimensionality of the data. However in recent years, several subtyping methods have been specifically developed for high dimensional data, which carries great potential for future applications [69, 70].

2.3.3 External validation

As for many scientific fields, a core challenge in fMRI is that the findings from most studies are not tested on external data, and for those that are, the findings often do not generalize well [71–73]. Even for supervised machine learning studies, where prediction performances are evaluated on 'out of sample' individuals, the performance often drop drastically when applied to external test datasets [17, 66]. E.g., in the review paper by Woo et al. they showed that the weighted mean accuracy of classification studies in psychosis was around 80% on the 'model development sample', while only few models were prospectively tested on external data. For those that tested on external data, the accuracy dropped substantially to a weighted mean of approximately 60% [17]. This is in contrast to neuroimaging studies in Alzheimer disease, where findings were

reproducible on external data with similar prediction performances. It should be noted that these findings are not specific for fMRI, but neuroimaging data in general.

The advantages of testing a model on external data is twofold. Firstly, it removes the'data-dependency bias' and thereby the risk of reporting a model that is overfitted to a specific dataset. Secondly, when only testing the final model on the external data, this further removes potential 'model flexibility bias' which can occur during model development, e.g., if several different machine learning models are trained [17, 74].

2.4 **Multi-site fMRI data**

Since acquisition of fMRI data is expensive (economically, time-wise and with respect to the burden that patients experience), most studies so far have relied on relatively small single site studies. However, in the last two decades an increased focus has been on multi-site datasets, which can be used to increase the sample size, speed up the data collection and which might even be necessary in cases where it is difficult to recruit participants. [2] Multi-site datasets exists in different formats. They can either come from coordinated initiatives, where MRI protocols, phenotypic assessments and inclusion/exclusion criteria have been standardized across acquisition sites. Examples of such coordinated initiatives include the 'Human Connectome Project' [75] and 'UK Biobank'[76], as well as multi-center clinical trials in drug development.

Alternatively, multi-site datasets can also come from data sharing initiatives such as 'OpenfMRI' which enables researcher to share their datasets with others.

Apart from increasing the sample size, multi-site data also includes more sources of heterogeneity, which can lead to site specific biases. These biases should be kept in mind when designing a study on multi-site data, particularly if the disease factors are confounded with the site bias. E.g. a extreme case would be if data from all patients were acquired at site A, while healthy controls from site B. In this case it would not be possible to determine if the found difference are related to the disease or reflect site biases between the two groups. Site related biases have been described in different ways in the literature, but we will use the description by Yamashiata et al. who have categorized multi-site biases into sampling and measurement biases [77, 78].

Measurement bias: This bias occurs due to technical factors that are different between the sites such as scanner manufacturer/types, image acquisition protocols, MRI coil use etc. Since measurement biases are a sources of non-biological variability, the goal is often to minimize the influence of these. This can be done in a number of ways as described in the next section.

Sampling bias: This kind of bias is related to biological differences in the population and clinical assessments between site. For example, different sites have access to to different patient populations, have different raters (and potentially rating systems) and might also differ in their inclusion/exclusion criteria. When the data is acquired during a coordinated initiative, the influence of the two latter examples can be reduced by standardizing protocols and assessment procedures across the sites.

---

[2]E.g. when a study is performed on a very narrow clinical population.

Both for single, and multi-site studies, it is important to consider how the inclusion/exclusion criteria are defined according to the question that is being studied. When using very strict criteria (e.g., only including young females with first psychosis, no comorbidity and before treatment onset), it will be difficult to recruit (often resulting in limited sample sizes) but the dataset would be quite homogeneous. A biomarker found in such a study might have a large effect size, but would likely generalize poorly to external data, since the biomarker was found on a small portion of the 'patient space' [21, 67]. On the contrary, a study with broad inclusion criteria the sample size will be much higher, however it might only be possible to find a limited number of biomarkers, since these have to be shared by the majority of participants. Furthermore, a biomarker from such a study might be non-specific, such as secondary effects related to differences in lifestyle [21].

2.4.1 Multi-site variability reduction

Measurement biases can be reduced through different steps all the way from data acquisition and preprocessing to how the predictive modelling is performed. In this section we have mainly focused on factors that influence the functional connectivity from multi-site studies, since this is the setting in which we used multi-site data.

Data acquisition: When multi-site studies are planned (e.g. in a multi-site research initiative of clinical trial), there are several factors that should be considered when attempting to standardize the acquisition parameters across sites. The multi-site standardization will often have to be a pragmatic choice, since most imaging sites sites are diverse in their equipment, which means that it will be difficult to include different sites if too high standardization requirements are set. A pragmatic choice can be to specify a recommended protocol, where some of the most critical factors are fixed (e.g. the magnetic field strength), while others can be implemented with more flexibility between sites. In 2009, Van Dijk et al. explored how the functional connectivity reliability was affected by different parameters in the acquisition (as well as preprocessing and other analytical procedures) across six sites [79]. Overall they found a moderate to high test-retest reliability, and that the correlation strength depended on the task (open eyes with fixation cross were best) and duration of the scan time (stabilized around 5 min), while other factors influenced the stability to a lower degree [79]. Similar results were also found in a study by Noble et al in 2019 who also found the largest reliability effect depending on the scan time [80], and it has been suggested that field maps (which allow for correction of distortion due to field inhomogeneity) should be included when possible [81]. Finally, in a review paper from Carmichael et al. they suggested that the following factors should be consistent across sites: type of pulse sequence, temporal resolution (TR), voxel size and slice thickness/spacing, number of observations, flip angle and field of view [18]. For studies that use data from retrospectively pooled data sharing initiatives (e.g.'OpenfMRI') these factors can be considered when choosing what datasets to pool. Whereas the parameters described above are related to technical factors, it should be noted that good data quality also highly depends on that the technician that acquires the data is sufficiently trained and that quality control is performed on the data.

Preprocessing: Preprocessing procedures aim to clean and standardize the data prior to further analyses, to minimize the effect of non-neuronal sources of variability. Preprocessing can

include many different steps, which has been shown to affect the final outcome [79, 80, 82, 83]. Overall, the degree of preprocessing can be considered in two ways. Firstly, a study can aim to tailor the preprocessing for the specific dataset to remove as many sources of non-neuronal variability as possible, the second approach is to keep the preprocessing at a minimal which can give increased robustness and generalizability between studies [84]. This is described in more details in the preprocessing section 3.1.

Multi-site harmonization (feature level): Even after standardizing imaging protocols and using preprocessing, earlier multi-site studies have shown that there are still site biases between the datasets [77, 85, 86]. This has motivated the development of multi-site harmonization methods. Different kinds of multi-site harmonization methods exists both with [77, 86] and without [85–88] travelling subjects. Multi-site harmonization methods and are further described in section 3.9.

Multi-site data in predictive modelling: Whereas the earlier methods attempt to remove the multi-site bias, it has also been suggested that this might not be needed if a sufficient amount of multi-site data is used to train the model. In an multi-site fMRI study on the Autism Brain Imaging Data Exchange database, Abraham et al. made a structured comparison on how different factors in the prediction pipeline, including the amount of data used for training (compared to testing) impacted the prediction results [89]. They performed their predictions either using 'inter-site' prediction, where data from a whole site was left out for testing, and 'intra-site', where a proportion of the participants from each site are used for the test dataset instead, as illustrated in Figure 2.4.
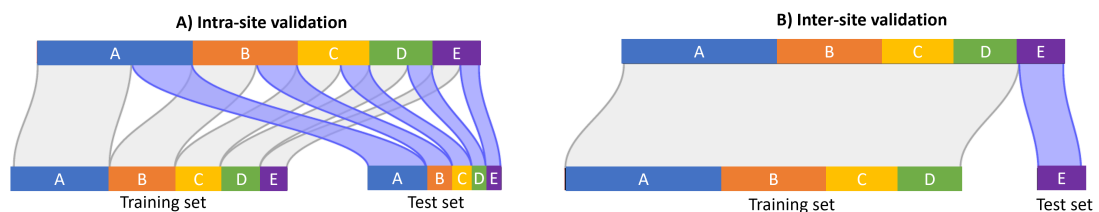
Figure 2.4: Intra-site and inter-site testing. In Intra-site cross validation (Panel A) data from all sites are used both in the training and testing set. In contrast, for inter-site cross validation (Panel B) data from one or more sites are left out as test data (external data). This figure is inspired by panel A of Figure 1 in Abraham et al. from 2017 [89].

Overall they found that the main difference between intra-site and inter-site site setting was an increased variability of the inter-site predictions, but that this disappeared when sufficient amount of data was used for training [89]. The influence of intra-site and inter-site predictions are further described and illustrated in section 4.2.

2.5 fMRI biomarkers – *where are we now*

The search for fMRI biomarkers in brain disorders such as schziophrenia has been ongoing for several decades. Whereas structural MRI biomarkers have already proven highly useful in many clinical applications, such as stroke [90] and multiple sclerosis [91], the only routinely used clinical application of fMRI is in presurgical mapping. Here, fMRI (often using a robust paradigm,

e.g. to determine language areas [92, 93]) is used to localize an area for surgery, mostly in patients with brain tumor or epilepsy [64]. A core challenge of clinical applications for other purposes, is that they mostly require fMRI to provide not only the location (as in presurgical mapping), but also the activation strength, for which there are still methodological challenges that need to be addressed to get reliable estimates on single subject level [64].

In the last decade, where the field has started to focus on data-driven methods and large multi-site datasets, there have been rapid advancement of multiple fronts [3], which give considerable reason to be optimistic about the future [15]. Overall, the field is still at stage of finding solutions for methodological challenges rather than contributing with specific and robust biomarkers for clinical practice [6, 94]. However, even though there are important challenges to be overcome, it has been argued that it is now a matter of when, rather then if, a solution to these is found [94–96].

2.5.1   fMRI biomarkers in drug development

The use of fMRI biomarkers in drug development is an example of where there is great potential, but which so far remains relatively limited due to a variety of biological, technical and strategic barriers [18]. In this PhD project we have not used data from any clinical trials, but the use of fMRI in drug development have been a strong motivation to look into the potentials and challenges related to multi-site imaging and reproducibility of fMRI biomarkers (Study 2-4). In this section we will thus give a short introduction to how fMRI can be used in drug development. For more comprehensive descriptions of the potentials and challenges, we refer to earlier reviews on the topic [18, 97, 98].

In **early phases (Phase 1 and 2)** of clinical drug development, fMRI can be used to detect the functional effect of a pharmacological intervention (pharmacodynamic biomarker). While positron emission tomography (PET) imaging is often used in drug development to measure the effect of the compound on a receptor level in the brain, it can not be directly used to measure multiple effects on more than one receptor type, nor target receptors for which no labelling has been developed. In contrast, fMRI can be used as a circuit engagement biomarker that reflects activity of multiple neurotransmitter systems on a brain circuit level [12, 18, 99].

In **later phases (phase 3 and 4)**, fMRI is more likely to be used to demonstrate normalization of a disease related fMRI signal or to give a more objective measure of disease modification, which can be used as (additional) evidence for a regulatory submission [18]. Since clinical trials in later phases include more patients, these trials are mostly carried out at multiple clinical sites which should be considered when planning a study.

Even though there are still methodological challenges to be solved, fMRI has been used in more than 1,000 clinical trials, and approximately one third of these had fMRI as the only primary outcome, as shown i a review by Sadraee et al. [100]. In this paper, they described fMRI as an outcome measure in clinical trails based on a systematic review in ClinicalTrials.gov from $1998 - 2018$ [100].

---

[3]As recently described by Calhoun et al. some of the fields with remarkable growth in recent years include deep learning, multi-modal fusion, and dynamic connectivity [15]

To obtain an overview of fMRI trials in schizophrenia, we performed a brief search on ClinicalTrials.gov and found 18 industry funded clinical trials in patients with schizophrenia (excluding trials that were terminated). Figure 2.5 shows the distribution of the trials in different phases (Panel A), the name of the industry sponsor (Panel B), and number of started trials per year (Panel C). First of all, we found that a wide range of companies have sponsored these trials, and that only few companies have funded more than one trial. Comparing the results of this short search with the review from Sadraee et al. we also found that the majority of trials where in phase 1, however whereas Sadaree et al. reporting an continuously increasing number of trials, we found that for schizophrenia only two trials have been started since 2016. When comparing the curve with the overall number of industry sponsored trials in schziophrenia, a similar trend is seen, where the number of trials peaked around year 2008, and then have steadily decreased thereafter [101]. This indicates that the decreased number of fMRI trials in schizophrenia are related to the elevated failure rate and number of pharmaceuticals companies that have stopped their activities in schizophrenia [12]. Other factors have likely also impacted this development such as the increased awareness about not only the promises but also challenges of fMRI as biomarker in drug development [18], lack of robust fMRI biomarker that are replicated across studies[6], and the COVID pandemic which have decreased the initiation of clinical trials [102].
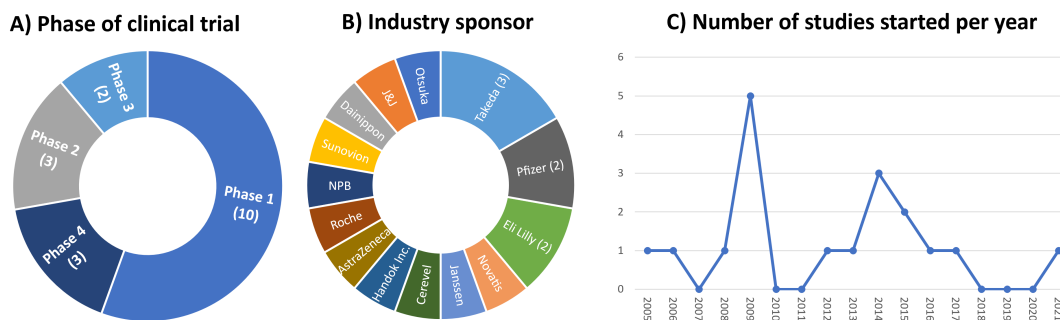
Figure 2.5: Overview of clinical trials in schizophrenia that used fMRI. A search on ClinicalTrials.gov for industry sponsored trials in patients with Schizophrenia that have used fMRI (performed on 6th of February 2023). This search included 18 studies, which here are listed according to their phase (Panel A), industry sponsor (Panel B), and number of trials started each year (Panel C). The full search history is added in appendix section, which also includes the full name of the industry sponsors A.4.

Examples of potential biomarkers

To round off, we here want to give two examples of potential fMRI biomarkers that could be used in drug development for patients with schizophrenia. Examples like these in part motivated our studies.

A reliable PANSS biomarker: Since the PANSS scale is often used as a endpoint in clinical trials, and is even considered the "gold standard" for assessment of antipsychotic treatment efficacy [103], a fMRI biomarker that reliably reflects changes in the PANSS scale would be very beneficial. E.g. such a biomarker could be included in an early clinical trial to investigate if the

treatment intervention would alter the activation. In this way, the biomarker readout could help to de-risk the clinical development plan, and even be used for regulatory support or even decision making. However, the latter would require very rigorous validation of the biomarkers analytical and clinical validity [7].

Patient stratification biomarker: Since schizophrenia is a very heterogenous disorder, treatments targeting a specific disease-mechanism may not produce the desired response in all patients. Here a fMRI biomarker that could be used to subtype patients into groups with a more homogeneous biology, which could be used stratify or enrich the population of patients included in a clinical trial.

2.5.2   fMRI biomarkers in Schizophrenia

Since fMRI can be used to study brain activation in the whole brain (including subcortical regions) and brain networks, it has become a popular tool for investigating brain alterations in schizophrenia. Abnormal brain activation have been linked to many different aspects of the disease, such as specific symptoms and different disease stages [24]; however, with diverging results which often have been attributed to the high heterogeneity of the disease and small sample sizes [21, 67]. Over the last two decades, an increasing number of fMRI studies have identified widespread functional connectivity changes in patients with SZ compared to healthy controls, supporting the view that schizophrenia can be characterized as a disorder of disorganized communication among brain networks[41, 104]. However, since the methods and results varies across studies, firm conclusions that can lead to clinically applied biomarkers, are yet to be made [6].

In the following sections we will give a short summary of the related literature for our four studies.

2.5.3   fMRI biomarkers for Schizotypy (Study 1)

The goal of risk prediction biomarkers it to identify individuals with increased risk of developing a disease, which provides the opportunity for early interventions and possibly even disease prevention. One way to identify risk prediction biomarkers of schizophrenia is to study the neurobiological underpinnings of schizotypy.

At the time when Study 1 was performed, only few studies had used fMRI in individuals with schizotypy either with conventional brain mapping approaches [105–109] or using predictive modelling [110–112]. Overall, these studies showed that various brain regions were altered in relation to schizotypy, and Modinos et al. had shown that their machine learning analysis had higher sensitivity compared to conventional brain mapping analysis [113].

In an a study from 2015 Wang et al. found positive correlations between the degree of social anhedonia and the activation in the temporal parietal junction and medial prefrontal cortex, during a social cognition fMRI task. Based on these findings, the **goal of Study 1,** was to build a machine learning pipeline with a broad selection of different feature extraction methods, to determine which of these were sufficiently strong to classify participants according to their degree

of social anhedonia (high or low). Furthermore, we aimed to validate the classification pipeline, including the use of decomposition methods for feature extraction through a task-paradigm classification for which an reliable gold standard for the class labels exists (e.g. theory of mind vs. physical control condition task).

2.5.4 Brain connectivity changes in Schizophrenia (Study 2)

Within the last five years, meta analyses and studies with relatively large sizes have consistently found that patients with SZ have hypoconnectivity in a most parts of the connectome [13, 89, 114–117], i.e., that the connectivity between brain regions are decreased in patients with schizophrenia compared to healthy controls. When focusing on data from **resting state fMRI**, alterations have been found for most resting state networks (RSN) [40]. These alterations are often categorized as either *within* a certain RSN e.g., connectivity between brain regions within the frontoparietal network (FPN), or *between* RSN connectivity constituting changes in connectivity between brain regions of different RSN, e.g., between the FPN and sensory motor cortex as illustrated in Figure 2.6.

In a meta study from Dong et al. in 2018, they found that schizophrenia was mainly characterized by hypoconnectivity *within* in the thalamus, DMN, affective, ventral attention, auditory and somatosensory network, as well as *between* several of the RSNs. In the latter, the ventral attention and FPN where "hubs" for connectivity alterations [13]. The only instance of hyperconnectivity was found between the affective and ventral attention network [13]. Furthermore, several earlier studies have found an imbalance between the DMN, FPN and salience network in patients with SZ, which supports the triple network model of psychopathology [40, 60, 117]. While these studies show interesting trends, robust conclusions have yet to be made, since the individual studies vary greatly in their methods and findings [13, 15, 40].

A) Within and between RSN connectivity **B) Connectivity matrix** **C) Somatomotor RSN**

Visual SoMo dATT vATT Limbic FPN DMN
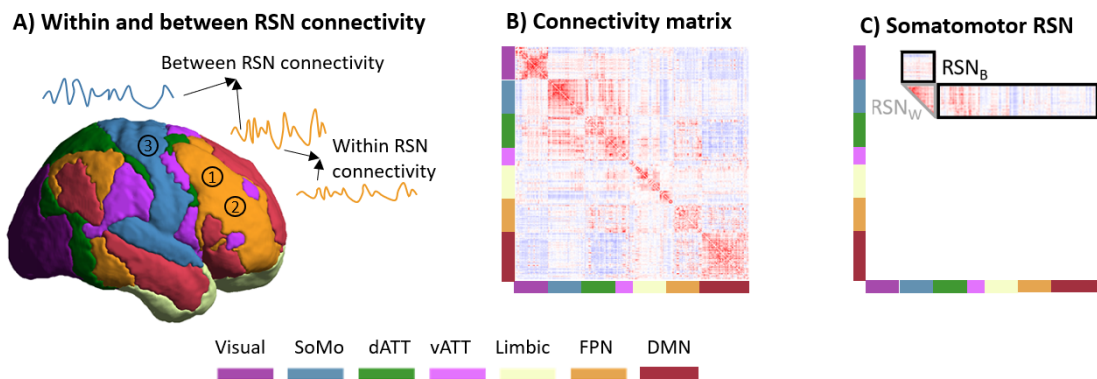
Figure 2.6: Illustration of within and between RSN connectivity. Panel A shows the 7-RSN parcellation from Yeo et al. [118], and an example of within (e.g. between ROI 1 and 2) and between (e.g. between ROI 1 and 3) RSN connectivity. Often parcellation based connectivity is shown in a connectivity matrix as visualized in Panel B. Panel C shows the interpretation of within (RSN_W) and between RSN (RSN_B) connectivity on a connectivity matrix. This figure is an adapted version of Figure 5 from Paper C.

In **goal of Study 2** was therefore to use one of the largest available multi-site datasets

with schizophrenia patients to classify the disease diagnosis and to test if the classification was reproducible on external data. To achieve this, we adjusted the prediction pipeline from Study 1 to have an increased focus on robustness and reproducibility. Furthermore, due to the missing consensus of earlier studies on which RSN is important to differentiate patients with SZ from healthy controls, we focused a substantial part of our analysis to determine if any individual RSN could drive the classification, and how these results related to weightmaps which are typically used to interpret the output from classification studies.

2.5.5   Prediction of PANSS scores (Study 3)

Since the PANSS scale is frequently used to measure the symptom severity of schizophrenia, several studies have attempted to find brain patters that correlate or even predict PANSS scores. This can be done by using the total PANSS score (overall measure of symptom severity), individual items of the scale, or the three PANSS subscales indicating the positive, negative and generalized dimension. The last two options can even be used to search for psychopathological subtypes in an attempt to disentangle the heterogeneity in schizophrenia [50].

Within the predictive modelling field, we know of four studies that used fMRI data to predict PANSS scores. The first was published by Koch et al. in 2015, who used fMRI data from the ventral striatum during a reward processing task and predict the total PANSS score, and found a high and significant correlation [119]. The other three studies used functional connectivity from resting state data to predict the PANSS positive and negative subscales. These studies found significant prediction performances when using an individualized connectivity extraction approach, while no significant results when using traditional group atlases [120–122]. Furthermore, to evaluate which brain regions were important for the predictions, they interpreted the weightmaps and found that it was mainly *between* RSN connectivity that drove the predictions. All of these studies were performed on single site data, and even though similar methods were used for the three last studies, they were applied on different clinical populations (patients with schizophrenia [120], adolescent-onset-schizophrenia [121] and first episode schziophrenia (before and after treatment with Risperidone) [122]). Furthermore, some studies have used brain activation correlations (instead of predictive modelling) to search for brain regions with a significant PANSS relation [123–125]. All in all, findings and methods of earlier studies are divergent and there is so far no fMRI biomarkers that have shown accurate and reproducible relations to the PANSS scale.

If it would be possible to find brain activation patterns that accurately and reliably predicted the PANSS score, this biomarker would be valuable both in drug development (as described in section 2.5.1) and to further understand how brain alterations are related to the symptom severity (PANSS total) and the to the three symptom dimensions specified by the subscales.

The **goal of Study 3** was to use regression-based predictive modelling to predict both the total PANSS score and the three subscales. This was done using the same features as in Study 2, such that multi-site data was used to train the model, and the reproducible was tested on external data.

2.5.6 fMRI based subtypes in Schizophrenia (Study 4)

Another way to tackle the internal heterogeneity of schizophrenia is to search for fMRI based subtypes. Compared to PANSS based psychopathological subtypes, these would not rely on any information from subjective measures such as diagnostic labels or clinical scales (e.g. PANSS), and they therefore have the potential to find more data-driven subtypes. Within schizophrenia, we know of three earlier studies that used fMRI for disease subtyping. The first studies from Brodersen et al. [126] and Yang et al. [127], demonstrated that clustering can indeed be used to subtype patients with SZ; however, these studies suffered from some methodological challenges [14] and their findings have not year been replicated. The third study was by Tokuda et al. from 2021 [128], where they seeked to identify a common brain network that could discriminate between different psychiatric disorders (including schizophrenia) and healthy controls. Since this study investigated "disorder differentiation networks" they did not aim to identify subtypes within disorders themselves.

The discovery of robust and reproducible subtypes (either from clinical scales or from fMRI data) would be highly beneficial to disentangle the high heterogeneity of schizophrenia, and possibly even between psychiatric disorders [128, 129]. Subtypes with a more homogeneous biology might even provide a natural basis for 'stratified psychiatry' [129] and can also have applications in drug development as described in section 2.5.1.

The goal of study 4 was to use multi-site data to search for fMRI based subtypes and to validate its reproducibility on external data. Since the fMRI based subtyping field is still at an exploratory state, a large part of our analyses were focused on determining the stability of the clustering solutions.

2.6 Overview of datasets

This section gives a overview of the datasets that were used for the four studies.

In study 1, (dataset D1), we included data from 76 college students from Guangzhou Medical University, where participants were selected such that they covered a continuous range of schizotypy. None of the participants had a history of drug abuse or psychiatric disorders. All MRI scans were acquired on a 3T Siemens Verio MR scanner at Guangzhou First People's Hospital in 2012 (single site study). We used fMRI data from a comic strip task that was designed to specifically probe theory of mind and empathy processing (two active task conditions). Furthermore, the comic strip task included two 'physical control conditions' which were designed to look similar to the social cognition tasks. The MRI specifications can be found in our publication of this study (Paper B). The study was approved by the Ethics Committee of the Institute of Psychology at the Chinese Academy of Sciences, and data collection was finished already prior to the work conducted in this thesis. The same dataset was also used in the master thesis of Lærke Gebser Krohne, where a range of analysis were performed which inspired the work of this study. However, all the analyses (apart from preprocessing) that are included in Study 1 and publications A and B, were performed after the completion of the master degree.

For study 2-4 we used multi-site data from two publicly available data bases: i) the Decoded Neurofeedback (DecNef) Project Brain Data Repository (DecNef), where we used dataset 3 'SRPBS Multidisorder MRI Dataset' as described in Tanaka et al. [19], ii) the Center of Biomedical Research Excellence (COBRE) datasets [20]. For all participants, we used the structural T1 weighted image, and 5-10 min resting state fMRI data with eyes open. More detailed information about the number of included participants and MRI acquisition parameters for each site can be found appendix table A.1. For the DecNef database, the data from some of the sites were acquired with a unified protocol [19, 77] (these sites are marked with an * in Supplementary table A.1), while data from other sites where not. Studies 2-4 were approved by the Institutional Ethical Review Board at the Technical University of Denmark, Department of Applied Mathematics and Computer Science (COMP-IRB-2022-03).

For each study we split the data into a discovery dataset which we used to train the different machine learning models, and an external test dataset to asses the reproducibility of our findings. The exact number of participants and sites included in each dataset differed between studies as specified in Table 2.1 and the following sections:

In study 2 (dataset D2) we aimed to maximize the amount of data, which means that we had an unbalanced dataset with more healthy controls (HC) than patients with schizophrenia. We constructed the splits between dataset such that approximately 70% of the data was used for training the models in the discovery dataset (D2a) and data from the remaining two sites were used in the independent test dataset (D2b), hence this splitting procedure served to assess between-site generalization.

In study 3 (dataset D3) we kept the same split between the discovery (D3a) and test dataset (D3b), but only included SZ patients that had a PANSS score available.

In study 4 (dataset D4) we created a balanced dataset in order to have the same amount of SZ patients and HC for for both the discovery dataset(D4a) and external test dataset (D4b, including the same two sites as in Study 2-3). The balanced dataset was constructed using the R package MatchIt [130] with nearest neighbor matching based on propensity scores (age and gender). We used exact matching on site and the quality of the matches were assessed through the balance of the covariates (age and sex) before and after matching (using diagnostic quantile-quantile plot (QQ) plots), and visual inspection of the propensity score distributions.

Further information and acknowledgement to the DecNef and COBRE datasets can be found in section A.3.1.

	Study 1	Study 2				Study 3		Study 4			
	D1	D2a		D2b		D3a	D3b	D4a		D4b	
	Healthy	HC	SZ	HC	SZ	SZ	SZ	HC	SZ	HC	SZ
$n_{participant}$	76	486	143	260	63	136	44	143	143	63	63
n_{sites}	1	8	3	2	2	3	1	3	3	2	2
Sex (♂/♀)	37/39	256/230	100/43	179/81	35/28	99/37	20/24	101/42	100/43	37/26	35/28
Age ($\mu \pm \sigma$)	19 ±1	40 ±16	36 ±12	34 ±12	42 ±10	36 ±12	42 ±10	35 ±10	36 ±12	42 ±11	42 ±10

Table 2.1: Demographics for the four studies Number of participants (healthy control(HC) and patients with schizophrenia (SZ)), number of sites, gender and age for the participants included in each of the four studies. For datasets D2-D4 (study 2 -4) the suffix "a" is used for the discovery dataset, and "b" for the external data.

FROM RAW FMRI DATA TO BRAIN FEATURES

3

This chapter describes the different steps that we used to move from raw data to interpretable brain features, which will be used for the subsequent analysis. The overall steps are illustrated in Figure 3.1 and include preprocessing, feature extraction, and multi-site harmonization methods. Feature extraction is an important step for fMRI data, due to the very high dimensionality of the raw data (in the order of 10-100.000). In this high dimensional feature space, the data lives on a limited manifold which only makes up a relatively small part of the space, hence dimensionality reduction can be used while retaining most relevant information [131]. If feature extraction is not used prior to prediction, the trained models are likely to be too specialized and generalize poorly to test data, which is known as overfitting. Furthermore, feature extraction is important to ease the the interpretation of the extracted features [17].

The aim of this chapter is to give a conceptual understanding of the different steps and how they have been used across the four studies of this PhD project. More detailed information about the specific implementations in each study are given in the corresponding papers for each study.
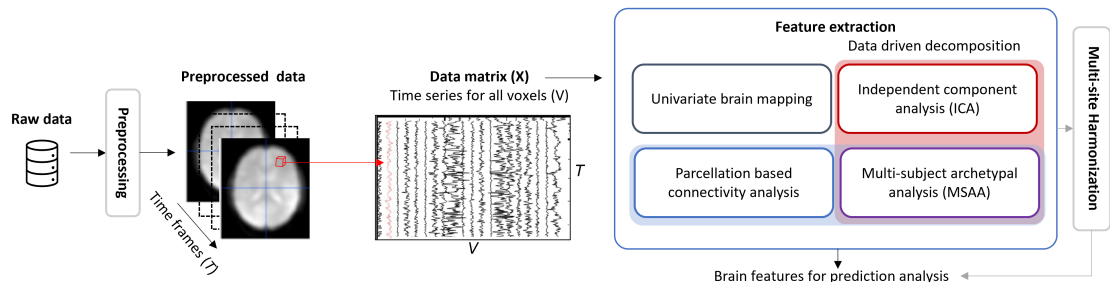
Figure 3.1: From raw fMRI data to brain features. Illustration of steps included in Chapter 3: i) preprocessing ii)feature extraction and iii) multi-site harmonization. From the preprocessed data, a data matrix $\mathbf{X}$ is constructed, which includes the time series for each voxel (V). This data matrix is then converted into interpretable brain features using four feature extraction methods. In a preliminary analysis, we have investigated the effect of applying multi-site harmonization as described in section 5.5

3.1 Preprocessing

When fMRI data is acquired, the BOLD signal is mixed with non-neuronal sources of variability, which can hamper the validity of inference and interpretability of the results [84, 132]. Therefore, most fMRI studies include preprocessing procedures to clean and standardize the data before statistical analysis. Preprocessing includes several steps, which can be split into two categories as described by Esteban et al. [84]. Firstly, preprocessed time series are derived from the original data after the use of signal corrections, spatio-temporal filtering and resampling onto a standard space

[1]. Secondly, nuisance signals and experimental confounds are estimated to effectively remove their contribution for further analysis.

Each preprocessing step include a range of different choices, which leads to a large number of parameters in these procedures. One may view these as free parameters to optimize the preprocessing procedures [133], however such an optimization may lead to overfitting, bias and can compromise comparison of findings across studies [82, 134]. Even though there have been several attempts to outline the best practices for preprocessing [83, 134], the large variability in data acquisition protocols has led to the use of customized preprocessing pipelines for most individual studies [135]. In 2019, Esteban et al. published 'fMRIprep', which is a data driven preprocessing pipeline for fMRI data that automatically adapts a preprocessing workflow for fMRI datasets with no or minimal manual intervention. fMRIprep has been developed to enable robust and reproducible preprocessing, and the authors have shown that it produces high quality results for fMRI data [84].

In Study 1, we used a customized preprocessing pipeline, while fMRIprep was used for Study $2 - 4$. The different preprocessing steps and parameter choices for each of our studies can be found corresponding papers.

3.2   Univariate brain mapping

For many years, the univariate brain mapping approach has been the traditional way to determine which brain regions are related to an outcome of interest. This is mostly done using a mass-univariate analysis where a parametric statistical test is performed for each voxel separately as illustrated in Figure 3.2.
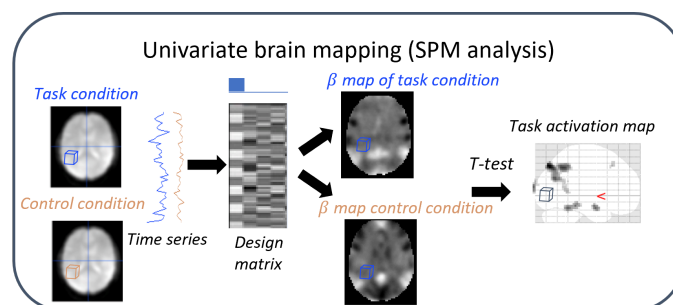
Figure 3.2: Illustration of Univariate brain mapping using Statistical Parametric Mapping (SPM). In a SPM analysis the following three steps are performed for each voxel. First, the time series are extracted from the preprocessed data. Secondly, the β parameters are estimated for both the task and the control condition, using the general linear model and a design matrix (which includes all explanatory variables (incl. task onset and duration)). Finally, to perform statistical inference on the parameter estimates, a parametric test (t-test) is performed to asses potential differences between the active and the control condition. This is repeated for all voxels, which results in a statistical parametric map to describe task activation.

We used mass-univariate brain mapping in Study 1, to determine the statistical parametric

---

[1]often using the the (Montreal Neurologic Institute (MNI) brain template, such that brain activation peaks can be reported in MNI coordinates

map between the task and control conditions. We did this using Statistical Parametric Mapping (SPM)[136] as described in Paper B. In a SPM analysis, the general linear model is used to described the observed data as a linear combination of exploratory variables and an error term. For each voxel v, the observed data $\mathbf{x} \in \mathbb{R}^{T \times 1}$ (where $t = 1, \dots T$ is the number of time frames, i.e. the time series for the given voxel) is given by

$$\mathbf{x_v} = \mathbf{D}\beta_v + \epsilon_v, \tag{3.1}$$

where $\mathbf{D} \in \mathbb{R}^{T \times p}$ is the design matrix including all p explanatory variables that are presumed to influence the data, $\beta_v \in \mathbb{R}^{p \times 1}$ is a parameter vector and $\epsilon_v \in \mathbb{R}^{T \times 1}$ is the noise term for each time series. It is generally assumed that ϵ_v is normal distributed and approaches such as pre-whitening are used to ensure that the elements are uncorrelated [137]. The parameter vector β_v is typically estimated by minimizing the sum of the squared residuals. Here eq. 3.1 is rewritten to the so called 'normal equation', by multiplying each site with the transposed design matrix ($\mathbf{D}^\top$) : $\mathbf{D}^\top \mathbf{x_v} = \mathbf{D}\mathbf{D}^\top \beta_v$. It can be shown that any β_v that satisfies the normal equation, will also minimize the sum of squares of the residuals $\hat{\epsilon_v}^\top \hat{\epsilon_v}$ [138]. Therefore the parameters can be estimated such that

$$\hat{\beta}_v = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{x_v}. \tag{3.2}$$

Statistical inference on the parameter estimates are then used to determine if there is a significant difference between the active and the control condition. This is performed using classical parametric testings, where the null hypothesis is that the task has no effect on the signal $H_0 : \mathbf{c}\beta = 0$ whereas the alternative hypothesis states the opposite $H_0 : \mathbf{c}\beta \neq 0$. Here $\mathbf{c}$ is the contrast, given as a linear combination of β_v. As the variance σ_v^2 is estimated from the data, the value below is distributed according to a Student's t-distribution under the null hypothesis

$$t_v = \frac{\mathbf{c}\hat{\beta}_v}{\sqrt{\mathbf{c}((\mathbf{D}^\top \mathbf{D})^{-1}\mathbf{c}^\top \hat{\sigma}_v^2}}. \tag{3.3}$$

Since this test is performed for each voxel, multiple comparison correction is needed. In the present study we used random field theory (RFT) which seeks to control the family-wise error while taking into account smoothness of the residuals, effectively making it less conservative than Bonferroni correction [139, 140]. The degree of smoothness includes both the intrinsic smoothness of the image acquisition process, and the additional smoothing applied during preprocessing. The intuition behind RFT is that the greater the smoothness the less severe is the multiple comparison problem is, and thus less stringent correction is needed [140].

3.3 Parcellation based connectivity analysis

Parcellation based connectivity analysis is a commonly used feature extraction method to estimate brain connectivity of fMRI data, and the main steps are illustrated in Figure 3.3. Different brain parcellation methods exists, and earlier studies have found that the selection of an adequate brain parcellation atlas is an important decision point for connectivity analysis[141]. However, so far

there is no 'gold standard', and parcellation methods vary greatly between studies [142, 143]. In this section, we shortly describe the theory of parcellation based connectivity methods, and then describe how we have used this feature extraction method across the four studies.
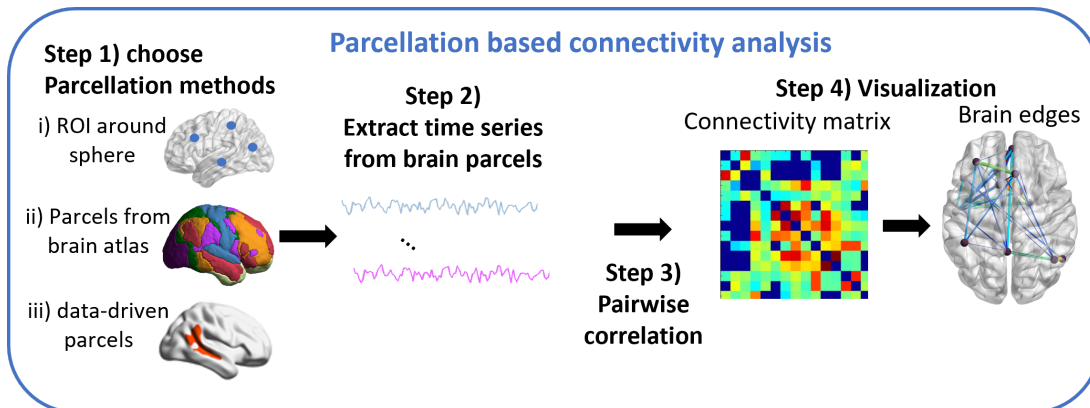
Figure 3.3: Illustration of Parcellation based connectivity analysis. Step 1, choose a brain parcellation, step 2: for each brain parcel (also called region of interest (ROI)), the time series is extracted. Step 3, the correlation between time series is calculated pairwise for all ROIs. Step 4, these correlations can be shown either in a connectivity symmetric matrix or as correlation strengths (edges) between ROIs on visualization of the brain.

The first step of a parcellation based analysis is to define a way to extract brain parcels (also often referred to as regions of interest (ROIs)). Overall brain parcels can be defined in different ways: i) a sphere around a center coordinate ii) parcels from a brain atlas or iii) using data-driven approaches to parcel the brain into different regions (Figure 3.3). The second step is to extract the time series from each brain parcel [2]. Thirdly, for each pair of brain parcels (for K parcels, this results in $(K(K-1))/2$ connectivity pairs), the correlation is estimated, e.g. this can be done using Pearson's correlation coefficient. Using the Fisher transform, the correlation coefficients are then transformed to their corresponding Z-values, such that the probability density function (PDF) of the distribution gets closer to the PDF of a normal distribution [144, 145]. Finally, the connectivity is visualized directly as a color coded connectivity matrix, or as correlation strengths (edges) between ROI on a visualization of the brain 3.3.

3.3.1 Brain atlases used in our studies

We have used parcellation based connectivity analysis in all four studies. In Studies $1-3$ we have used the 'sphere approach', where brain parcels were defined as a sphere around a set of center coordinates, and in Study 4 we have used a brain atlas.

Parcellations using in Study 1

In Study 1, we used two different parcellations, where the center coordinates came from either a literature study or a pooled condition analysis as described below. The ROIs were defined as all voxels in a sphere with a radius of 8mm.

---

[2]This which can be done either by using the mean signal or first eigenvariate which reflects the strongest single source across the included voxels of the brain parcel [138]

Literature study coordinates: the first ROI definition included 25 center coordinates which were determined through a literature study for the social cognition task (comic strip task of theory of mind and empathy processing) that was used in the study. For each ROI, we took into account both the specificity of the presented task [146–148] and the consistency across studies [149, 150]. The center coordinates are illustrated in Panel A of Figure A.1.

Pooled condition analysis coordinates: in an attempt to use a more datadriven approach to find center coordinates, we also extracted center coordinates from SPM analysis of the given task. Here we used a 'pooled condition analysis' which reflects the pooled effect of both of the social tasks conditions compared the the controls. This analysis resulted in 16 cluster, of which the coordinates are shown in Panel B of Figure A.1.

Parcellation used in Study 2 and 3

In Study 2 and 3 we used a center coordinate based brain atlas presented by Seitzman et al [151]. This atlas is an extension to the 'Power atlas' from Power et al [152] which included 264 ROIs. In the Seitzman atlas rsfMRI data was used to get an improved representation of ROIs in the subcortex and cerebellum [151], which resulted in a '300ROI' atlas, that can be downloaded from the Greene lab website. We excluded 25 ROIs since these were outside the the field of view for most participants in our studies. For the remaining 275 ROIs (hencefort referred to as '275ROI atlas'), we assigned a RSN label to each coordinate, using the 7-network parcellation from Yeo et al. [118], since this was the RSN parcellation that we have used for the decomposition methods in these studies.

Parcellations used in Study 4

In Study 4, we used two different atlases: i) the '275ROI atlas' as described above, and ii) the 'Allen atlas' which is an ICA based atlas based on resting state data from 603 healthy controls which was presented by Allen et al in 2011 [153]. The atlas includes 28 components that are each assigned to one of the following RSNs: basal ganglia, auditory, sensory motor, visual, default mode, attentional and frontal, as specified by Allen et al. [153].

3.4 Decomposition methods

Decomposition methods are instances of unsupervised machine learning algorithms that aim to decompose the observed data as an outer product of matrices, often imposing a specific structure and/or sparsity on these [154]. Given a dataset $\mathbf{X} \in \mathbb{R}^{T \times V}$ with T time frames and V voxels, the goal of linear decomposition method are to find a basis set $\mathbf{A} = a_1, \ldots, a_K$ such that the linear space spanned by $\mathbf{A}$ is a close reconstruction of the data $\mathbf{X}$

$$\mathbf{X_i} = \epsilon_i + \sum_{k=1}^{K} a_k s_i(k) \tag{3.4}$$

where the data from each subject $\mathbf{X_i}$ is characterized by unique coefficients $s_i \in \mathbb{R}^K$ for the basis set of $\mathbf{A}$ and a residual noise term ϵ_i. Typically the goal is to find a solution such that $K << V$,

i.e, that the dimension of the basis set is much lower that the number of voxels, which gives rise to the dimensionality reduction. In matrix notation, eq. 3.4 can be written as $\mathbf{X} \approx \mathbf{AS}$ where $\mathbf{S} = s_1, \ldots, s_n$.

Decomposition methods such as independent component analysis (ICA) (described below) are a model-free (do not rely on any brain atlas) alternative to parcellation based connectivity analysis, which simultaneously explore connectivity across the whole brain, either on a individual or group level [154].

We have used the decomposition methods ICA and multi-subject archetypal analysis (MSAA) in Studies $1 - 3$. Both methods result in a set of subject specific spatial maps, which reflect brain networks and corresponding time series. These two methods are described in more detail below, followed by a description of how we have labelled the spatial maps (section 3.7), and investigated consistency of decomposition solutions across datasets (section 3.8).

3.5   Independent component analysis (ICA)

Independent component analysis (ICA) is one of the most frequently used data driven methods to derive brain networks from fMRI data [155, 156]. In most neuroimaging applications spatial ICA (illustrated in Figure 3.4) is used to find spatial maps that are maximally independent according to a sparsity promoting prior distribution. Hence, the components are largely non-overlapping, and the timeseries within each of the components will typically have high temporal correlation.
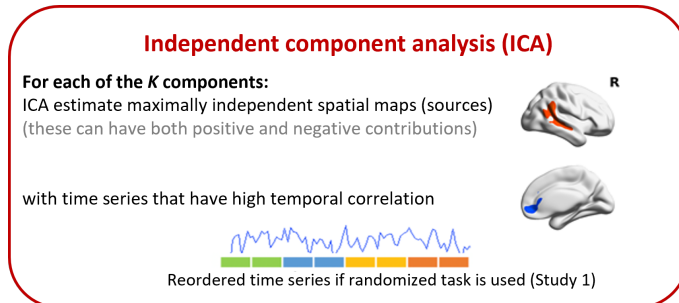
Figure 3.4: Illustration of Independent component analysis (ICA). For each of the K components (specified by the user), ICA estimates a set of maximally independent sources ($\mathbf{S}$) and corresponding time series which have a high temporal correlation (included in the mixing matrix $\mathbf{A}$).

The classical (noise free) ICA decomposition model can be defined as

$$\mathbf{X} = \mathbf{AS}, \tag{3.5}$$

where $\mathbf{S} \in \mathbb{R}^{T \times V}$ is the sources matrix. Here, each row represents a statistically independent map, and $\mathbf{A} \in \mathbb{R}^{T \times T}$ is the mixing matrix, which is formed by the time series of each component as columns in the matrix. ICA aims to estimate an unmixing matrix $W = A^{-1}$ such that

$$\mathbf{Y} = \mathbf{WX}. \tag{3.6}$$

is a good approximation to the real sources, $\mathbf{S}$.

Group ICA Group ICA (GICA) is a commonly used method for fMRI analysis of group data, which precedes ICA with a data compression steps using principle component analysis (PCA).

First PCA is applied to the data of each participant to reduce the dimensionality and hence computational complexity, then these datasets are temporally concatenated, and finally group PCA further reduces the temporal dimension, such that only K principle components are remaining. Spatial ICA is then applied to decompose the compressed matrix $\mathbf{X}$ into spatially independent spatial maps as given in equation 3.5. Finally, the participant specific spatial source and corresponding time series are obtained by using dual regression [157] (used in Study 2 and 3) or back-reconstruction (used in Study 1) [158].

Prior to the ICA analysis, a number of components (K) needs to be specified. In all studies, we have used the minimum description length (MDL) criteria, which is a model selection method that penalizes the likelihood with respect to the number of comments for a given model [159, 160]. Throughout Study $1-3$, we have applied ICA using the GroupICATv4.0a GIFT toolbox[161]. We have used the 'Infomax' algorithm, which separates the mixture of independent sources by maximizing the mutual information which the output $\mathbf{Y}$ contains about its input $\mathbf{X}$, using a fixed Sigmoid non-linearly function [161, 162]. This algorithm produces a result which is equivalent to maximizing the likelihood of the sources given an assumed independent source distribution.

3.6 Multi-subject archetypal analysis (MSAA)

Archetypal analysis is a decomposition method similar to ICA which includes additional constraints aimed towards easing the interpretation of the features [163, 164]. As for ICA the aim is to identify a low rank representation $\mathbf{X} \approx \mathbf{AS}$, where $\mathbf{A}$ includes the archetypes, which are given by K distinct temporal profiles (analogous to the mixing matrix of ICA), $\mathbf{S}$ is the archetypal source matrix (spatial maps), which reflect the contribution of the different archetypes to each observation in the data matrix $\mathbf{X}$. The archetype matrix $\mathbf{A}$ is defined as $\mathbf{A} = \mathbf{XC}$ where $\mathbf{C}$ is the seed region matrix, that specifies how the archetypes are generate from the data $\mathbf{X}$.
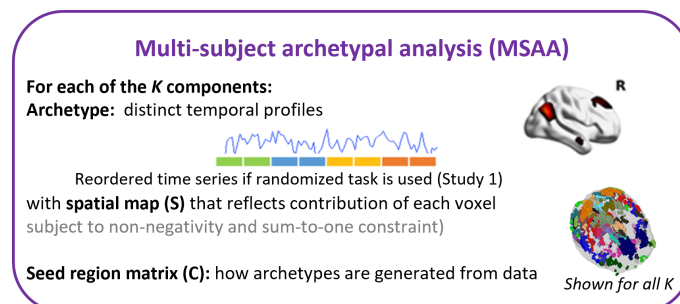
Figure 3.5: Illustration of Multi-subject archetypal analysis (MSAA). For each of the K components (specified by the user), MSAA finds a set of characteristic archetypes (time series, columns in $\mathbf{A}$) and their corresponding spatial maps ($\mathbf{S}$), which reflect the fractional contribution of each voxel to that specific archetype. The seed matrix ($\mathbf{C}$) includes the voxels that generate the archetype. Here the seeds for all K archetypes are shown. This process is shown in more detail in Figure 3.6 and 3.7 for the whole brain and spotlight MSAA respectively.

The components are found by minimizing the sum of squares (Frobenius norm) reconstruction

error subject to a set of convex (non-negativity and sum-to-one) constraints

$$\underset{\mathbf{C},\mathbf{S}}{\operatorname{argmin}}\|\mathbf{X}-\mathbf{X}\mathbf{C}\mathbf{S}\|_F^2 \tag{3.7}$$

$$\mathbf{C},\mathbf{S} \quad \geq \quad 0, \quad |\boldsymbol{c}_k|_1 \quad = 1 \quad \forall k, |\boldsymbol{s}_v|_1 \quad = 1 \quad \forall v.$$

The constraints on $\mathbf{C}$ enforce that the archetypes are a convex combination of the data $\mathbf{X}$, while they on $\mathbf{S}$ ensure that each observation in $\mathbf{X}$ is reconstructed by a convex combination of $\mathbf{X}$.

Multi-subject archetypal analysis (MSAA) is an extension of archetypal analysis for group analysis of fMRI data, which was presented by Hinrich et al. in 2016 [165]. Since MSAA is a relatively recent decomposition method compared to ICA, this section includes more details for the component estimation. In MSAA, the low rank representation for a subject i, is given as $\mathbf{X}^{(i)} \approx \mathbf{X}^{(i)}\mathbf{C}\mathbf{S}$ such that each subject have subject specific spatial maps $\mathbf{S}^{(i)}$ and archetypes ($\mathbf{A}^{(i)} = \mathbf{X}^{(i)}\mathbf{C}$). On the contrary, the seed region matrix $\mathbf{C}$ is the same for all subjects, which ensures the consistency across subjects.
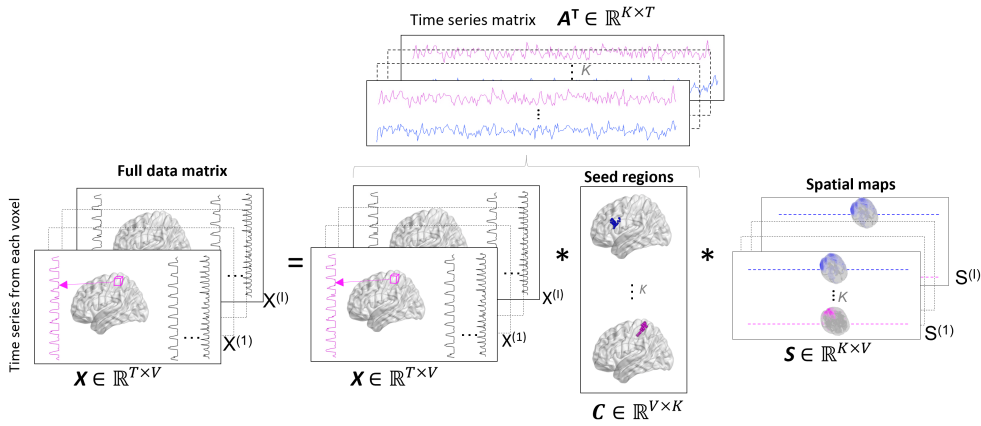
Figure 3.6: Illustration of Multi-subject archetypal analysis (MSAA). Schematic overview of spatial MSAA. For each subject $i = 1,\ldots, I$, the data matrix $\mathbf{X}^{(I)}$ includes the time series ($t = 1,\ldots, T$) for all voxels ($v = 1,\ldots, V$). Through alternating least squares optimization, MSAA determines the common seed region matrix $\mathbf{C}$, as well as a set of K temporal ($\mathbf{A}^{(i)} = \mathbf{X}^{(i)}\mathbf{C}$) and spatial ($\mathbf{S}^{(i)}$) components for each of the I subjects. This figure is a slightly adapted version of Figure 2 in Paper B [166]

MSAA enables heteroscedastic noise modelling, to overcome the challenge of inter-subject variability which arises both due to changes in the brain activation (neuronal function and cerebrovascular response) and noise sources (residual confounds that are not controlled for, such as psychological noise or movement). This is done by including an additional noise term ($\sigma_{i,v}$), which can capture both voxel (v) and subject (i) specific noise.

In MSAA, the linear model for each subject can be formulated as:

$$\mathbf{X}^{(i)} \approx \mathbf{X}^{(i)}\mathbf{C}\mathbf{S} + \mathbf{E}^{(i)}. \tag{3.8}$$

Where the noise ($\mathbf{E}^{(i)}$ with columns $\epsilon_{i,v}$) is assumed to be independently distributed with a Gaussian distribution such that $\epsilon_{i,v} \sim \mathbf{N}(0, I_T\sigma_{i,v}^2)$, where $I_T \in \mathbb{R}^{T \times T}$ is the identity matrix. In

MSAA, the components are found by minimizing the negative log-likelihood ($\mathcal{L}$) given by:

$$-\log(\mathcal{L}) = \sum_{i=1}^{I} \sum_{v=1}^{V} \frac{T_i}{2} \log\left(2\pi\sigma_{i,v}^2\right) + \frac{\|\mathbf{x}_v^{(i)} - \tilde{\mathbf{X}}^{(i)}\mathbf{C}\mathbf{s}_v^{(i)}\|_F^2}{2\sigma_{i,v}} \tag{3.9}$$

Since the MSAA model includes the subject and voxel specific variances and mixing matrix, the optimization is given as

$$\underset{\mathbf{C};\mathbf{S}^1,\mathbf{S}^2,\dots,\mathbf{S}^I;\sigma_1^2,\sigma_2^2,\dots,\sigma_I^2}{\operatorname{argmin}} -\log(\mathcal{L}) \tag{3.10}$$

$$\begin{aligned} \mathbf{C} &\geq 0, & |\mathbf{c}_k|_1 &= 1 & \forall k \\ \mathbf{S}^{(i)} &\geq 0, & |\mathbf{s}_v^{(i)}|_1 &= 1 & \forall v, k \end{aligned}$$

As for other decomposition methods, estimating the components ($\mathbf{C}$ and $\mathbf{S}^{(i)}$) jointly leads to a non-convex optimization problem [163, 165], and the solution is therefore found by alternating optimization using projected gradient descent. Here, the gradient is calculated in the projected space, holding one parameter fixed while optimizing the other (and vice versa), such that each sub-optimization becomes a convex optimization problem. To do so, the matrices $\mathbf{S}^{(i)}$ and $\mathbf{C}$ are rewritten to their ℓ_1 normalization invariant variables as suggested by Eggert and Körner in 2004 [167]:

$$\tilde{s}_{k,n}^{(i)} = \frac{s_{k,n}^{(i)}}{\sum_{i'} s_{k',n}^{(i)}}, \text{ and } \tilde{c}_{n,k} = \frac{c_{n,k}}{\sum_{n'} c_{n',k}} \tag{3.11}$$

with $v = 1,.,\dots V$. The gradients of the subject-specific archetypal mixing matrices $\tilde{\mathbf{S}}^{(i)}$ are independent for each subject and can therefore be updated independently, while the gradient of the common seed matrix $\tilde{\mathbf{C}}$ is influenced by all subjects. The gradients are given by the following two equations

$$G^{\tilde{\mathbf{S}}^{(i)}} = 2\left(\tilde{\mathbf{C}}^\top\tilde{\mathbf{X}}^{(i)}\right)^\top\left(\tilde{\mathbf{X}}^{(i)}\tilde{\mathbf{C}}\tilde{\mathbf{S}}^{(i)} - \tilde{\mathbf{X}}^{(i)}\right)\operatorname{diag}(\boldsymbol{\sigma}_i^{-2}), \tag{3.12}$$

and

$$G^{\tilde{\mathbf{C}}} = 2\left(\sum_{i=1}^{I}\tilde{\mathbf{X}}^{(i)}\right)^\top\left(\tilde{\mathbf{X}}^{(i)}\tilde{\mathbf{C}}\tilde{\mathbf{S}}^{(i)} - \tilde{\mathbf{X}}^{(i)}\right)\left(\tilde{\mathbf{S}}^{(i)}\operatorname{diag}(\boldsymbol{\sigma}_i^{-2})\right)^\top. \tag{3.13}$$

To find the updates of $\mathbf{S}^{(i)}$ and $\mathbf{C}$ (in the normal space) the chain rule is used, resulting in the updates

$$s_{k,v}^{(i)} \leftarrow \max\{\tilde{s}_{k,v}^{(i)} - \mu_v^{\mathbf{S}^{(i)}}(g_{k,v}^{\tilde{\mathbf{S}}^{(i)}} - \sum_{k'} g_{k',v}^{\tilde{\mathbf{S}}^{(i)}}\tilde{s}_{k',v}^{(i)}), 0\}, \tag{3.14}$$

and

$$c_{k,v}^{(i)} \leftarrow \max\{\tilde{c}_{k,v}^{(i)} - \mu_v^{\mathbf{C}}(g_{k,v}^{\tilde{\mathbf{C}}} - \sum_{v'} g_{v',k}^{\tilde{\mathbf{C}}}\tilde{c}_{v',k}), 0\} \tag{3.15}$$

Given the ℓ_1 normalization invariant variables for $\mathbf{C}$ and $\mathbf{S}^{(i)}$ the latent variables $\sigma_{i,v}^2$ have a closed form solution and can therefore be solved directly, differentiating the log-likelihood function in Eq. 3.9 with respect to $\sigma_{i,v}^2$

$$\frac{\partial \log(\mathcal{L})}{\partial \sigma_{i,v}^2} = \frac{T_i}{2\sigma_{i,v}^2} - \frac{\|\mathbf{x}_v^{(i)} - \tilde{\mathbf{X}}^{(i)}\mathbf{C}\mathbf{s}_v^{(i)}\|_F^2}{2\sigma_{i,v}^4}. \tag{3.16}$$

Equalizing Eq. 3.16 to zero and isolating $\sigma_{i,v}^2$ results in

$$\sigma_{i,v}^2 = \frac{\|\mathbf{x}_v^{(i)} - \tilde{\mathbf{X}}^{(i)}\mathbf{C}\mathbf{s}_v^{(i)}\|_F^2}{T_i}. \tag{3.17}$$

As seen in Eq. 3.9 when $\sigma_{i,v}^2$ tends towards zero, the negative log-likelihood will approach infinity. Therefore, to ensure the numerical stability, a minimal value for $\sigma_{i,v}^2$ is

$$\frac{2\|\mathbf{x}_v^{(i)} - \tilde{\mathbf{X}}^{(i)}\mathbf{C}\mathbf{s}_v^{(i)}\|_F^2}{T_i} \geq \sigma_{i,v}^2. \tag{3.18}$$

Since the joint estimation of parameters in MSAA is a non-convex problem, there is a risk that the optimization identifies a local rather than a global minimum, which means that different initialization (i.e., initial assignment of the components which are being estimated ($\mathbf{S}^{(i)}$ and $\mathbf{C}$)) can lead to different solutions. A common way to tackle this, is to repeat the runs with different initializations, and select the solution that maximizes the log likelihood given in Eq. 3.9. [163].

As for ICA, the user needs to specify the number of components, *K*. In our studies we used the same *K* as for the ICA analysis, which was determined using the MDL criteria [159]. This is expected to be a good approximation since the expressiveness of the models are largely equivalent disregarding the constraints.

In the description of MSAA so far, the components are estimated using the full data matrix $\mathbf{X}$ in equation 3.8. We will also refer to this approach as 'whole brain' MSAA (as introduced by Hinrich et al. [165]), which is in contrast to 'spotlight MSAA' where only a subset of the columns (voxels) of the data $\mathbf{X}$ are used to optimize the seed region matrix.

3.6.1   Spotlight MSAA

Spotlight MSAA can be a useful approach to search for archetypes that are related to a certain set of brain regions. I.e., if the data comes from a task or patient population where there is a hypothesis that brain region A plays an important role, then spotlight MSAA can be used to enforce to algorithm to find archetypes with a seed in that region. The overall concepts and mathematical derivations for spotlight MSAA were already given by Hinrich et al. in 2016 [165] but the first application and investigation of its stability was performed during Study 1 or this thesis.

For spotlight MSAA the subject specific model (equation 3.8 for wholebrain MSAA) is given as:

$$\mathbf{X}^{(i)} \approx \tilde{\mathbf{X}}^{(i)}\mathbf{C}\mathbf{S} + \mathbf{E}^{(i)}, \tag{3.19}$$

where $\tilde{\mathbf{X}}^{(i)}$ indicates the subset of voxels (*M*) in which the seed region matrix can find archetypes. This is specified by a 'spotlight mask' which is given as an input to the algorithm as illustrated in Figure 3.7.

The spotlight MSAA can be viewed as a bridged version of a data-driven decomposition and a parcellation based connectivity analysis. On one hand, it is 'less data-driven' since the seed regions are restricted within a user-defined spotlight mask. On the other, it can enable the algorithm to find archetypes that are related to a certain brain region of interest which might
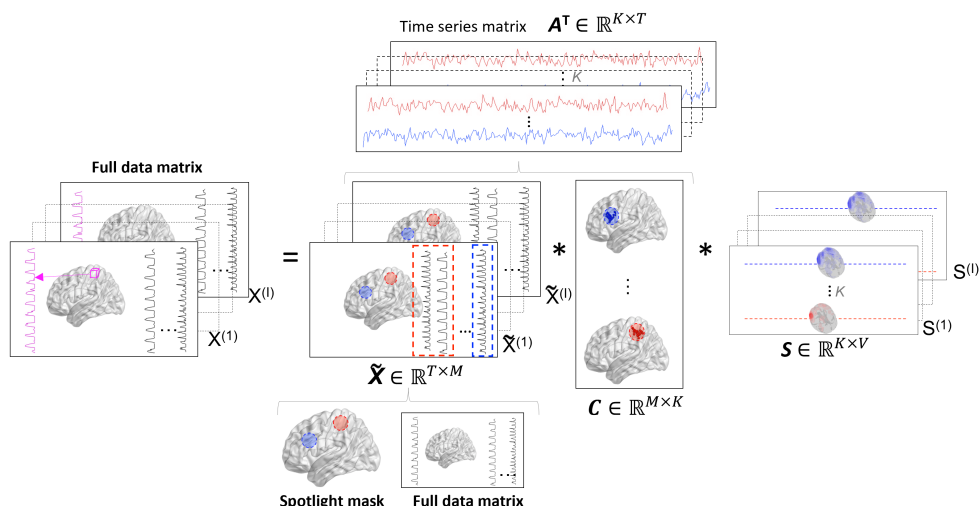
Figure 3.7: Illustration of spotlight Multi-subject archetypal analysis (MSAA). Schematic overview of spatial MSAA. For each subject $i = 1,\ldots, I$ the data matrix $\mathbf{X}^{(I)}$ includes the time series ($t = 1,\ldots, T$) for all voxels within the spotlight mask ($m = 1,\ldots, M$). The spotlight mask is given as an input to the algorithm, to specify in what brain regions MSAA can find archetypes. This figure is a slightly adapted version of Figure 3 in Paper B [166].

otherwise not have been extracted due to the presence of other 'sources' that are more strongly expressed in the data. E.g., if the study is focused on social cognition (as in our Study 1), then the spotlight approach can be used to extract archetypes related to brain regions that are believed to be involved in social cognition. These might otherwise not be found if they are less strongly expressed than other non-task specific networks (e.g. sensory motor networks). Compared to conventional parcellation based approaches, the spotlight MSAA still searches for the optimal seed (**C**) within the spotlight mask, instead of using the average signal of the region.

3.7 Interpretation of decomposition components

The output of each decomposition method is a set of subject specific spatial maps and time series for each component, which can be used for subsequent prediction. The interpretation of the components are usually based on the brain regions that they include.

In Study 1, we have used both the time series and spatial maps for classification. We described the networks by the regions that were included in the spatial networks. In study 2, where we used resting state data, we defined the networks according to which resting state network (RSN) it reassembled the most. There is no gold standard for how to do this, and in many studies the 'network label' is assigned based on visual interpretation. Others have assigned components to a previously defined RSN parcellation using the spatial correlation of networks [168, 169]. We opted for the latter approach, using the 7-RSN parcellation from Yeo et al [118] as described in Paper C to make the assignment more objective.

3.8 Transfer learning

Data-driven feature extraction methods such as ICA and MSAA find the brain networks or regions that best explain the dataset, which can be both a benefit but may also generate challenges. On the one hand, the extracted features represent strong trends for the specific datasets and the networks can be more sensitive compared to parcellation based approaches where the atlas dictate certain structure. On the contrary, the extracted features can be overfitted to the data on which they are extracted, which would lead to lower generalizability between studies. Furthermore, there is no guarantee that the same networks are found again when rerunning the analysis on another dataset (or even the same dataset given that decomposition is in general non-convex). When brain networks from decomposition methods are used for subsequent predictions, this poses a challenge when testing the reproducibility on external data. One could argue that if brain networks are "truly reliable" they should reproduce across datasets. In practice, this is partly true, e.g. characteristic RSNs such as the DMN are found repeatedly across datasets, but there will also be substantial differences between datasets. This means that it will not be exactly the same brain networks that are found across datasets, even if the same task is performed. Variability between datasets can arise both due to sampling and measurement biases as described in section 2.4. In Study 2, we therefore looked into three different transfer learning approaches to overcome this challenge. We use the term 'transfer learning' to indicate that information from one decomposition is transferred to the next, similarly to what was done in Cai et al. [170]. In these approaches that we investigated, the transfer learning was restricted to the 'feature level' and did not include any interventions on the predictions which were performed later on. This is somewhat different from how the term transfer learning is also used in the machine learning field, where it can also refer to methods that can be used to re-purpose a machine learning model that is trained on one task to another, related task [171]

For both the decomposition methods, we investigated the following three transfer learning approaches to find feature across the discovery dataset (D2a) and external test dataset (D2b).

Approach 1 Here we ran the decomposition analysis separately on the external test (D2b) dataset. The only "transfer learning" information for this approach is that we used the same number of components and algorithm settings as for D2a.

Approach 2: Here we ran the decomposition analysis on the merged dataset (D2a + D2b). In this way, the decomposition was performed on data from all participants, which in a way breaks the independence between the discovery and the test dataset. This will be described in more detail in section 4.1.1. However, since the decomposition is not informed about the phenotypic label, this does not bias the prediction performance as such.

Approach 3: Here we directly used the output from the discovery dataset to perform the decomposition on the external test dataset. For ICA this was done by using the same dual regression procedure that was also used to create subject specific spatial maps [157]. I.e., for each participant in external data (D2b), we used dual regression with the ICA decomposition map **S** from the discovery data. For MSAA, this was done by keeping the common seed matrix **C** from the discovery dataset, and then we performed a few additional MSAA iterations (until

convergence) to optimize the subject specific spatial maps **S** and heteroscedastic noise estimations ($\sigma^2_{i,v}$) for each participant in D2b.

Please note that these transfer learning approaches refer to how we found features on the external test dataset. When training the models on the discovery data, we used features that were found only on the discovery dataset.

3.9   Multi-site harmonization on feature level

A way to reduce multi-site bias is to apply multi-site harmonization on the feature level. This kind of bias removal has been inspired from the field of genomics, where harmonization methods have been developed to remove so called 'batch-effects' which arise when processing multi-site genomics data at laboratories with different equipment and at different times. One popular method to remove batch effects i called ComBat, which was developed in 2007 by Johnson et al. [172]. In recent years, ComBat has been extended to other fields, including MRI where it has been adapted to structural MRI [173, 174] and fMRI connectivity data [85, 86].

We used ComBat harmonization for a preliminary analysis on parcellation based functional connectivity features, as described in section 5.5. Here we used the Matlab based implementation of ComBat which is available at GitHub (https://github.com/Jfortin1/ComBatHarmonization).

Effect size of site effect

To quantify the 'severity' of the multi-site bias, we used Cohen's d (CD) to measure the site related effect size, similar to earlier multi-site studies [86, 175]. More specifically, for each connectivity feature, we calculated the effect size of how different the connectivity from site i was from the mean connectivity of the remaining sites. Cohen's D was then calculated according to

$$\text{CD} = \frac{|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|}{\sigma_p}. \tag{3.20}$$

Where $\bar{\mathbf{x}}_1$ includes the connectivity features from the participants included in site i, $\bar{\mathbf{x}}_2$ includes data from the participants from the remaining sites, and σ_p is the pooled standard deviation across all sites. This was repeated for all connectivity features and all sites (leaving out data from one site each time).

ComBat harmonization of FC features

ComBat is a multi-site harmonization tool which uses empirical Bayesian estimation to estimate and remove multi site variability. For a dataset **X** which is here a multiway-array including functional connectivity data from site $j = 1, \dots, J$, subject $i = 1, \dots, I$ and connectivity feature $k = 1, \dots, K$, the ComBat model can be written as

$$\mathbf{X} = \alpha + \mathbf{D}^\top \beta + \gamma + \delta \epsilon. \tag{3.21}$$

Where $\alpha \in \mathbb{R}^{K \times 1}$ is the mean connectivity value (across sites and subjects), $\mathbf{D}^\top \in \mathbb{R}^{J \times I}$ is the design matrix (including covariates of interest which should stay in the data), and $\beta \in \mathbb{R}^{K \times 1}$ are the

corresponding regression coefficients. Finally $\gamma \in \mathbb{R}^{J \times K}$ and $\delta \in \mathbb{R}^{J \times K}$ are terms for the additive and multiplicative site effect for each connectivity feature.

The goal of ComBat is thus to simultaneously estimate the biological and non-biological effect, and then to remove these estimated additive and multiplicative site-related effects from the data. The ComBat harmonized data is then given as

$$\mathbf{X}^{\text{ComBat}} = \frac{\mathbf{X} - \hat{\alpha} - \mathbf{D}\hat{\beta} - \gamma^*}{\delta^*} + \hat{\alpha} + \mathbf{D}\hat{\beta} \tag{3.22}$$

where the asterisk symbol * indicates that these are Empirical Bayes estimates. For more detailed descriptions of the method we refer to earlier publication of ComBat on MRI data [86, 173].

PREDICTIVE MODELLING AND SUBTYPING

4

This chapter gives an introduction to the machine learning methods that were used for clinical predictions and disease subtyping throughout the studies. In Studies $1-3$ supervised machine learning was used to predict a phenotypic measure (either classification using a binary group membership or a continuous outcome), while we in Study 4 used a unsupervised clustering algorithm to search for more homogeneous disease subtypes. The overall structure, including aims and machine learning methods for each study are summarized in Figure 4.1.
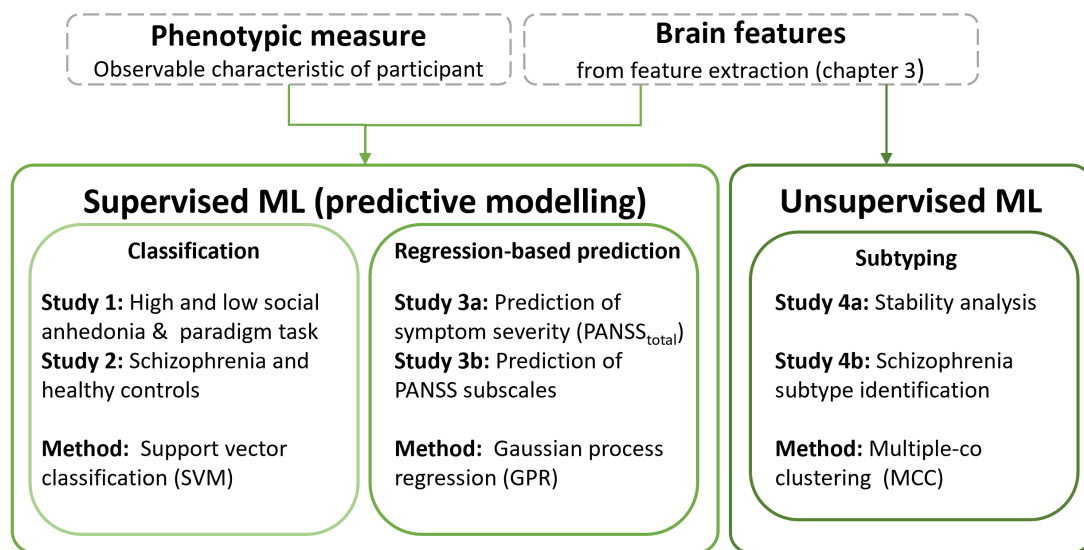
Figure 4.1: Overview of methods for predictive-modelling and disease subtyping. In Studies $1-3$ we explored ways of using supervised machine learning (ML) to predict phenotypic measures, while we in Study 4 aimed to search for data-driven disease subtypes using clustering (unsupervised ML method).

4.1 Neuroimaging based predictions

The goal of predictive modelling is to predict individual differences in phenotypes using e.g. neuroimaging (in our case fMRI) data. Figure 4.2 illustrates the overall steps involved in training and evaluating a prediction model. Step 1: the data (brain features and phenotypic measure for each participant) is split into a training and a test dataset. Step 2: On the training data, the prediction algorithm selects the most relevant features, and summarizes these to produce a model (mathematical function) which maps the high dimensional neuroimaging data onto the low dimensional phenotypic measures. Step 3: the model is applied on the test data to predict the phenotypic measure. Step 4: the model performance is evaluated by comparing the observed and the predicted phenotypic measures. Since the performance of prediction models is tested on data that was not used to train the model, they have the potential to uncover more robust

35

When using cross validation this process is repeated K times
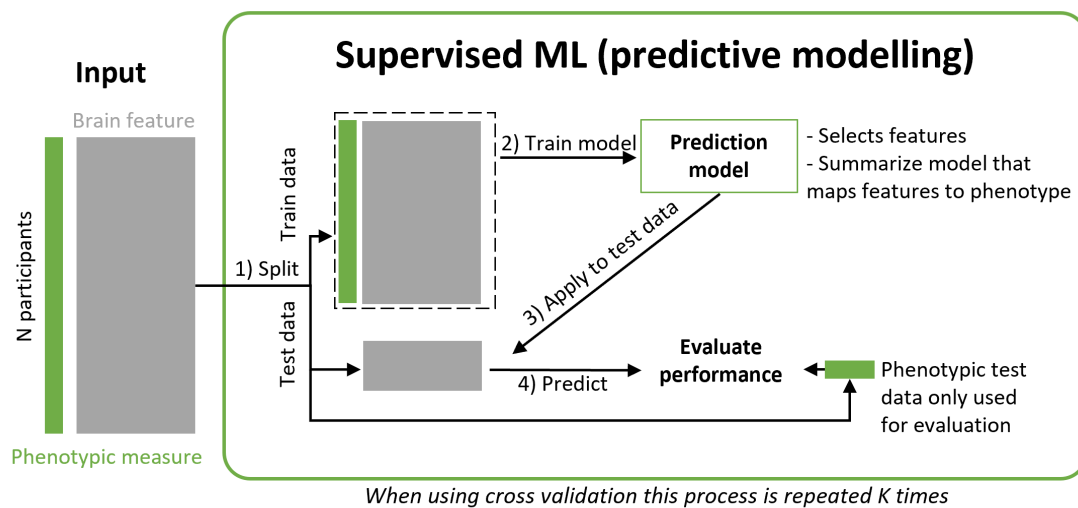
Figure 4.2: Introduction to supervised machine learning. Four overall steps of predictive modelling. In step 1, the data is split into a training and test dataset. In step 2 the model is trained, and then in step 3, applied on the test dataset, for which the performance is evaluated (step 4).

biomarkers. However, as the last decade of predictive modelling has shown, these models also have their limitations, and important considerations must be taken into account [15, 74].

4.1.1 Important considerations for cross validation

Cross validation is a internal validation strategy, where the process of splitting the data (step 1) is repeated multiple times for different combinations of the training and testing data. This is commonly implemented using K-fold cross validation, where the dataset is randomly divided into K non-overlapping subsets of equal size. The model is then trained on $K-1$ subsets and tested on the remaining subset. This process is repeated K times, leaving a new subset out each time. The choice of K affects the performance of the prediction. When K is large, much data is used for training, which generally will improve the performance, but since less data is left for testing, the variance will also increase [74]. Choosing K is hence a tradeoff between bias and variance.

There are two 'special cases' of K-fold cross validation which are commonly referred to as 'leave-one-out cross validation' (K = number of participants), and 'split half cross validation' (K = 2). Leave-one-out cross validation has often been used in studies with small sample sizes to maximize that data that is available for training, but it has been shown that the combination of this validation strategy and small sample sizes can lead to overfitting and should therefore performed with care [176]. For larger sample sizes (more than 200 participants), K = 5 or 10 have shown to be a good compromise between variance and bias [177, 178].

Stratified cross validation: Whereas the cross validation strategies described above presumes random datasplits between the training and test data, it might be important to ensure a balance of important characteristics between the dataset, which can be done using stratification. For example, if the aim of the prediction is to classify the group membership, then stratified cross validation can be applied to ensure that each cross validation subset has the same proportion of

participants in each group. Furthermore, stratification can also be used to ensure a balance of confounding factors, such as age and gender.

Nested cross validation: When a prediction model includes hyper-parameters that needs to be tuned during the training phase, this should be performed using nested cross validation to ultimately validate only a single model on the test dataset [176]. Here the data is divided into three subsets: a training, validation and test dataset. The hyper-parameters are tuned by running an repeated 'inner-loop' cross validation using the first two datasets. Then the model selected from the inner loop is validated on the outer loop with the test data. Participants may be reassigned to subsets repeatedly to assess the robustness of model selection and performance.

4.1.2 Independence

: Since the performance of predictive models is evaluated on data that was not used to train the model, it is often stated that the model is tested on unseen data, in other words that the training and test datasets are independent. One example where this can become challenging is during feature extraction which is typically done on the whole dataset prior to the prediction analysis, as illustrated in Figure 4.1. It should therefore be carefully considered if, and how this influences the independence of the training and test dataset. E.g. if a decomposition method, such as ICA, is used to extract brain networks on the entire dataset, this means that the networks of the training and testing are no longer completely independent. Since the decomposition did not include any information about the phenotypic measures (which are later used for prediction) this 'dependence' does not directly bias the prediction analysis but it should still be considered to what degree this could influence the results. A suggested solution to this problem would be to run the decomposition *within* each cross validation step [74]. Whereas this would overcome the break of independence, it comes with other quite complex challenges. Firstly, re-estimating the brain networks in each cross validation split, has a high computation burden, which can be severe limitation if the dataset is very large (e.g. for multi-site datasets). Secondly, if the decomposition is rerun, the same brain networks are not necessarily found between different cross validation splits, which would severely hamper the interpretability of the prediction model. This is why we have looked into different transfer learning approaches in Study 2 as described in section 3.8.

4.2 Testing models on external data

Even though cross validation has so far been the most frequently applied validation strategy, testing the models on external dataset (independent dataset of which no data was used for training the model) offers stronger evidence of the models reproducibility. When a model reproduces on external data, it shows that model does not only fit the specific dataset, but provides evidence that the model also generalizes to a more heterogenous group (more sampling and measurement bias) which indicates that it could be representative for the general population.

The importance of having external data which is not at all used for model development has been highlighted by many review papers, as well as in 'prediction competitions'. In the latter, a part of the data is shared with the participants for training a model, and the winner will be

chosen based on the model which achieves the best performance on an external dataset which is only made available after the model training has ended. Such studies have shown that it is rarely the best performing models from the training data which generalize to the external dataset, which indicates that the best performing models are overfitted on the training data[179, 180].

The benefit of testing models on external data are twofold. Testing the model on independent samples, eliminates the 'data-dependency bias', and when only testing the final model on the new data this further removes potential 'model flexibility bias' which can occur during model development, e.g., if several different machine learning models are trained [17, 74]. Developing models that reproduce on external data is therefore a crucial step towards moving biomarkers from research to clinical practice [15, 17].

When **multi-site data** is used for predictive modelling, it is strongly recommended to leave data from one or more sites for external validation. Such a prediction setup is called inter-site validation. However, sometimes this might not be feasible, and it can even give rise to other challenges. E.g. if the site factor is confounded with the phenotypic measure, inter-site validation can become misleading. In such cases it might be necessary to perform 'intra-site' validation. In this setting a proportion of the participants from each site are used for the test dataset as described in section 2.4.1 and Figure 2.4.

In Study 2 and 3 of this PhD project, we have used stratified intra-site cross validation to train models on a multi-site 'discovery dataset', while keeping data from two sites as external data, which was used to test the reproducibility of the models.

4.3   Classification of group membership

So far, most predictive modelling studies in the field have focused on group membership classification, where neuroimaging features are used to predict a binary phenotypic measure, e.g. a diagnostic label. An advantage of binary classification is that it gives an intuitive outcome, which can be compared with traditional group level analysis. However, if the heterogeneity within each group is large, or the phenotypic measure is actually reflecting a continuum, binary classification might not be very informative.

In this PhD project, we used binary classification in the first two studies. The goal of Study 1 was to classify participants with high and low social anhedonia, and the goal of Study 2 was to classify patients with SZ and healthy controls. In both studies we used Support vector Machines (SVM) to perform the classification analysis.

4.3.1   **Support vector machines**

The support vector machine (SVM) is one of the most frequently applied methods for classification of fMRI data [181, 182]. For binary classification, the goal of SVM is to find a decision boundary that separates the data from the two groups as illustrated in Figure 4.3. To minimize the risk of overfitting, the decision boundary is chosen such that it maximizes the margin, which is the perpendicular distance between the nearest data points in each class and the boundary [183, 184]. This is illustrated in Panel A of Figure 4.3, where each data point is a participant, and the shape

(circle and triangle) indicates the class membership. For illustrative purposes this example uses a two dimensional feature space, but the concepts readily generalizes to higher dimensional settings.
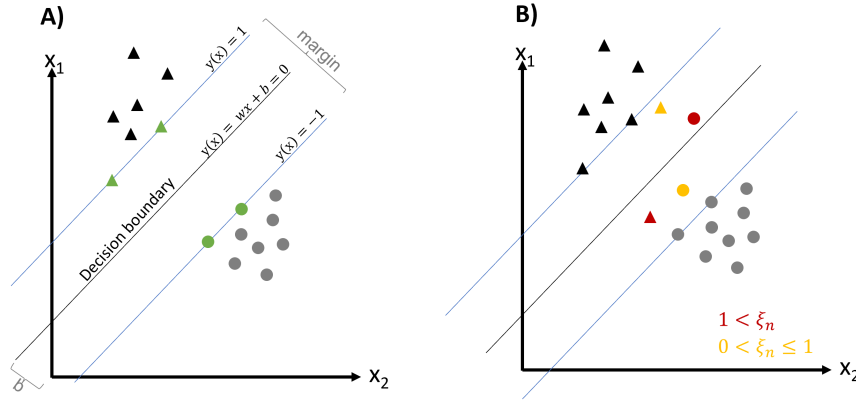
Figure 4.3: Overview of Support vector machine (SVM). Panel A illustrates the overall concept of a binary SVM classification in two dimensions. Here each data point represents a participant and the shape (triangle and circle) indicates the group membership. The grey lines show the maximum margin hyperplane between the samples of the two classes and b is the bias (offset from origin). Panel B illustrates the soft margin SVM, where misclassifications are allowed by introducing a slack variable ξ_n. Here $0 < \xi_n \leq 1$ for points which are inside the margin and correctly classified, whereas $\xi_n > 1$ for points that are wrongly classified.

For the classification of data from a new participant $\mathbf{x}^*$, the discriminant function $\mathbf{y(x)}$ is given as

$$y(\mathbf{x}^*) = \mathbf{w}\mathbf{X}^\top \phi \mathbf{x}^* + b. \tag{4.1}$$

Here $\mathbf{X}$ is the training dataset, $\mathbf{w}$ is a weight vector, b the bias (offset from origin) and ϕ denotes the feature-space transformation (kernel). The kernel can be used to transform the data, in cases where it is not possible to find a linear decision boundary that separate the two groups. The rationale is, that if the data is not separable in the current dimension, then adding dimensions can make them separable [185].

Given a training dataset of I participants $\mathbf{x}_1, \ldots \mathbf{x}_I$ and their corresponding phenotypic measures $t_1, \ldots, t_I$ which indicate a binary group assignment (-1 for group A and 1 for group B). A new participant $\mathbf{x}^*$ will be classified according to the sign of $y(\mathbf{x}^*)$, i.e., that participant is classified as being in group A if $y(\mathbf{x}^*) < 0$.

The solution for $\mathbf{w}$ and b, such that the margin is maximized can be found by solving the optimization problem

$$\min_{\mathbf{w},b} \frac{1}{2} \|\mathbf{w}\|^2. \tag{4.2}$$

This is solved subject to a set of linear inequality constraints which ensure that all participants are classified correctly. The equivalent dual problem can be obtained by using Lagrange multipliers, allowing the use of efficient algorithms for inference. It can be shown that according to the Karush-Kuhn-Tucker conditions, only datapoints (participants) that lie on the maximum margin of the decision boundary form the basis for the classification model. These participants are called

support vectors, marked with green in Panel A of Figure 4.3. This makes SVM a sparse kernel classifier which reduces the computational complexity [186].

Soft margin SVM: In descriptions above, it is required that all participants are classified correctly. In practice, this is not always possible, nor preferable, since a model which separates the training data perfectly might be overfitted and therefore may not generalize well to other datasets. Soft margin SVM is a method which aims to overcome this by allowing the model to misclassify some participants in the training dataset. This is done by introducing slack variables ξ_n, which are $0 < \xi_n \leq 1$ for points which are inside the margin and correctly classified, whereas $\xi_n > 1$ for points that are wrongly classified, illustrated in Panel B of Figure 4.3.

Adding ξ_n to eq. 4.2 the optimization function becomes

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \|\mathbf{w}\|^2 \tag{4.3}$$

where the parameter C controls the trade-off between the slack (cost of misclassification) and the width of the margin.

When a dataset is unbalanced (more participants in one group than the other), additional weights can be used to penalize the class with more participants such that equation 4.3 becomes

$$C_B \sum_{i \in I_B} \xi_B + C_A \sum_{i \in I_A} \xi_A + \frac{1}{2} \|\mathbf{w}\|^2, \tag{4.4}$$

where I_A and I_B indicate the participants for each group respectively. The soft-margin constants (C) are typically chosen according to the ratio between the two classes $\dfrac{C_A}{C_B} = \dfrac{n_A}{n_B}$, which is equivalent to upsampling the under-represented class such that sample sizes would be balanced.

4.4 Regression-based prediction

Regression-based prediction is a supervised machine learning method where the goal is to predict a continuous outcome, instead of a binary group membership as for the classification. It holds a great potential to provide more detailed understanding of phenotypic measures which are better represented by a continuum rather than a group membership, such as symptom severity. However, compared to binary classification, regression-based prediction can be more challenging because it aims to quantitatively estimate the specific score of a continuous measure over a range of that variable instead of 'just' determining the group membership [187].

In the Study 3 we used regression-based modelling to predict the symptom severity (PANSS$_{\text{total}}$) and PANSS subscales in an attempt to address the internal heterogeneity of schizophrenia. For this, we used Gaussian process regression as described below.

4.4.1 Gaussian Process regression

Gaussian process regression (GPR) is a non-parametric, Bayesian approach to regression where the model can provide uncertainty estimates and can learn the noise and smoothness parameters from the training data. Unlike other supervised machine methods, such as SVM which aim to

learn an exact value for each parameter in the function, the Bayesian approach infers a probability distribution of all values.

Given a input data matrix $\mathbf{X}$ with I participants and d features and a phenotypic target vector $\mathbf{y}$ which specifies the phenotypic measure for each participant, the goal of a GPR is to learn a function from the training data such that a new target y^* can be predicted given data from a new participant $\mathbf{x}^*$.

A Gaussian process (GP) can be uniquely defined by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ such that $\mathcal{GP} \sim \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ similarly to how a Gaussian distribution is given by its mean and covariance. The mean function is typically a constant [188, 189], and the covariance function could be any functional which takes two input arguments such that $k(\mathbf{x}, \mathbf{x}')$ generates a non-negative definite covariance matrix $\mathbf{K}$. The output can be thought of as a measure of similarity between the two arguments. One of the most frequently used covariance functions is the squared exponential, which commonly is referred to as radial basis function (RBF):

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2} |\mathbf{x} - \mathbf{x}'|^2\right). \tag{4.5}$$

This function has the two parameters lengthscale (ℓ) which determines the smoothness of the model, and magnitude (σ_f^2) which determines the distance from the function to its mean. This means that $k(\mathbf{x}, \mathbf{x}')$ tends towards σ^2 if the inputs are close to each other, and decreases exponentially as the squared distance between the inputs increases.

There are two alternate and equivalent perspectives of GP models: **Function space view:** Here the GP is seen as a distribution over functions $f(x)$ where the distribution is directly used to model the data. Predictions are made by placing a zero-mean GP prior over the functions and subsequently using Bayes rule to find the posterior distribution evaluated on the training data.

Weight space view: Here it is more straight forward to see how GPR is a Bayesian extension to a traditional linear regression problem, which is given by

$$y = f(\mathbf{x}) + \epsilon = \mathbf{x}^\mathsf{T} \mathbf{w} + \epsilon \tag{4.6}$$

where $\mathbf{w}$ is a weights vector and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ is a Gaussian noise term. For a GP model, prediction is made by placing a zero-mean GP prior over the weights and then computing the posterior distribution as

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \theta) = \frac{p(\mathbf{w}|\theta) p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \theta)}{p(\mathbf{y}|\mathbf{X}, \theta)} \tag{4.7}$$

Where θ is a vector of hyperparameters, $p(\mathbf{w}|\theta)$ is the prior, $p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \theta)$ denotes the likelihood and $p(\mathbf{y}|\mathbf{X}, \theta)$ is the marginal likelihood. The marginal likelihood (also called model evidence) can be written as $p(\mathbf{y}|\mathbf{X}, \theta) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$.

The prediction of a new test observation $\mathbf{x}^*$ is found by integrating over all possible values for $\mathbf{w}$ weighted by their posterior probability such that

$$p(f^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*, \theta) = \int p(f^*|\mathbf{w}, \mathbf{x}*, \theta) p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \theta) d\mathbf{w}. \tag{4.8}$$

In this way, GP predictions are a weighted average of all possible linear models, given the prior assumptions and training data $\mathbf{X}$.

Figure 4.4 illustrates an example of regression-based prediction with a traditional regression model, such as SVM (Panel A) and GPR (Panel B). Whereas the first returns the function that fits the training data the best, GPR (Panel B) provides a mean prediction function and a certainty for each input **x**. It can be seen that the model has a high certainty for **x** close to training points, whereas it becomes more uncertain in regions where training data was not available.
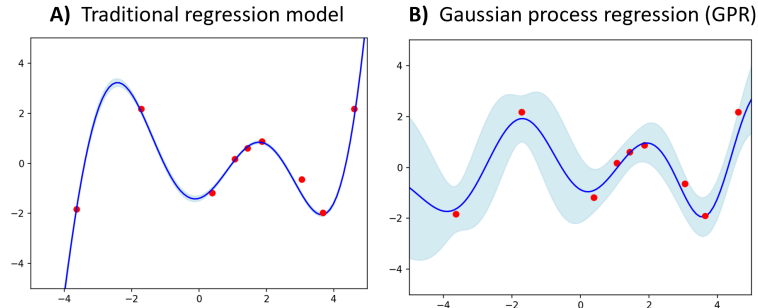
Figure 4.4: Gaussian Process regression (GPR) example. Panel A shows the result from a traditional regression model, which outputs the single function (blue line) that fits the datapoints (red dots) the best. Panel B shows the result from a GPR, which includes both a mean prediction function (blue line) that represents the most likely output, and a certainty estimate (blue shaded area).

Since the likelihood and prior of a GPR model are GPs the posterior predictive distribution is also Gaussian and can thus be computed in closed form

$$p(f^*|\mathbf{X},\mathbf{y},\mathbf{w},\mathbf{x}^*,\theta) \sim \mathcal{N}(\mu,\sigma^2) \tag{4.9}$$

where $\mu = \mathbf{k}^{*\top}(\mathbf{K}+\sigma_n^2\mathbb{I})^{-1}\mathbf{y}$ and $\sigma^2 = k(\mathbf{x}^*,\mathbf{x}^*) - \mathbf{k}^{*\top}(\mathbf{K}+\sigma_n^2\mathbb{I})^{-1}\mathbf{K}^*$

Here, **K** is a kernel matrix that describes the covariance between each observation, such that $\mathbf{K}_{i,j} = k(\mathbf{x_i},\mathbf{x_j},\theta)$, and $\mathbf{k}^*$ is a vector with the covariance between the training data **X** and new test observation $\mathbf{x}^*$.

4.5   Disease subtyping using fMRI

As introduced in earlier sections, the aim of fMRI based disease subtyping is to search for biologically defined subgroups of patients which have a more homogeneous biology. Even though disease subtyping on neuroimaging data has been a goal for several decades, the field is still at an exploratory state, where most studies are at a 'proof of concept' level and have not yet been integrated in any clinical practices [14]. Mostly subtypes of patients are identified using different unsupervised machine learning methods for clustering aiming to partition observations (here participants) in a dataset such that those within the same group (referred to as a cluster) are more similar to each other than those in other groups. Several different clustering algorithms exists, which differ in terms of how they evaluate and handle within and between cluster similarity [154].

As for predictive modelling, there are some inherent challenges in clustering to identify disease subtypes. Firstly, clustering is an ill-defined problem, which means that there is neither an

unique well-defined solution nor definition of what a cluster is [190]. Secondly, many clustering algorithms requires the user to define the number of clusters prior to the analysis, which in most cases is not a trivial task and can influence the outcome of the clustering to a large degree. Finally, as for predictive modelling the generalizability is of great importance since the whole point of unveiling a disease subtype from a given dataset is to extend these results to a broader subset of the general population. If a clustering solution (set of subtypes) is only valid within the given dataset, the model is overfitted to the data it has been trained on, and does not have much clinical utility [14, 70].

The high dimensionality of fMRI data has been a challenge for earlier subtyping studies, since the high number of features compared to observations (participants), enhances the challenges described above. However, in recent years, an increasing number of clustering algorithms have been developed specifically for high dimensional data, which holds a great potential for the field [69]. One way to mitigate the high dimensionality of fMRI data is to use polytopic learning methods, where clustering is performed on several datatypes (e.g. on data from fMRI and clinical scales). Here, combining fMRI data with data from a clinical scale can help to extract neurobiological information about disease related trends that might otherwise have been 'overlooked' [14, 191]. Another potential method is subspace clustering, where the clustering solution only needs to be present in a part of the feature space, which can ease the discrimination between group and also make it more reliable[69].

In Study 4, we used a multiple co-clustering method that is based on Bayesian mixture models, and which has shown promising results in earlier studies [191, 192]. Our main motivation for using this method is the ability to identify several different clustering solutions, and thereby disease subtypes, on the same dataset (a kind of multi-view clustering). In the next section, we will describe the multiple co-clustering algorithm, and specify how we have used it to search for subtypes that are related to the schizophrenia diagnosis.

4.5.1 Multiple co clustering

Multiple co-clustering (MCC) is a polytopic learning method, that can deal with different types of data, such as connectivity features and phenotypic measures, which are modelled with different distributions. Furthermore, the algorithm can handle missing values, and infers the numbers of clusters in a data driven way, such that they do not need to be specified by the user. The overall idea is that MCC partitions the features into several groups (called views), and within these views subject and feature clustering is performed. The algorithm simultaneously partitions the data such that:

- **views**: features with similar subject clustering are assigned to different views
  For each view, the data is further partitioned into:
- **feature clusters**: where features with similar distribution are grouped
- **subject clusters**: where subjects are grouped into subject clusters

Figure 4.5 illustrates and example of this, here the input data comprises of four numerical and four binary features from five subject (Panel A). The solution of the MCC algorithm separates this data into two views (Panel B) (this can be considered as a feature selection step), which makes it

possible to obtain different subject clustering solutions (one of each view) for the same dataset. Within each view, additional subject and feature clustering is performed to bundle observations with similar distributions. E.g. in this example, the first three features are included in View 1, which has three subject clusters (indicated by vertical colorbar to the left) and two feature clusters (indicated by horizontal colorbar on the top). View 2 includes the remaining five features, which are co-clustered into two subject clusters and three feature clusters. As shown in View 2, both numerical (feature 3) and categorical (feature 4-8) features can be in the same view, but not in the same feature clusters.
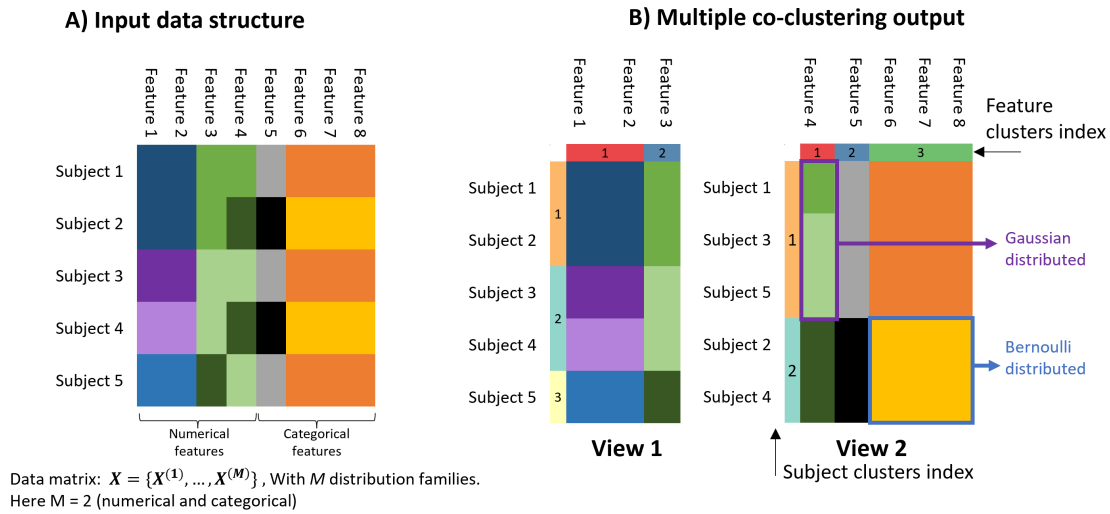
Figure 4.5: Illustration of multiple co-clustering (MCC). Panel A shows the input data which contains five subjects with four numerical and four categorical features. The output of the MCC algorithm (Panel B) includes two views, which each include further subject and feature clusters. The vertical colorbars (to the left of each view) indicate the subject cluster assignment, e.g. View 1 has three subject clusters. The horizontal colorbar (top) indicates the feature clusters, e.g. two feature clusters for View 1. This figure is inspired by Figure 1 from an earlier publication of the MCC method [192]

4.5.2   Model explanation

MCC is based on non-parametric Bayesian Mixture models, where each view can include features that are modeled with different distributions. The current implementation of the MCC algorithm can include Gaussian, Poisson and Bernoulli distributions, depending on the type of the underlying data [191]. The distribution of each feature needs to be specified by the user such that the data matrix **X** consists of M specified distribution families such that $X = X^{(1)}, \ldots, X^{(M)}$ where $m = 1, \ldots M$.

Overall, the MCC algorithm uses a hierarchical Dirichlet process to model a set of feature and subject partitioning tensors (which indicate the subject and feature cluster assignment), and uses variational inference to optimize the hyperparameters of the model. Finally, a univariate distribution is fitted for each cluster using conjugate priors.

Generative model for feature and subject partitioning

The MCC models a feature partition tensor **Y** and a subject-partition tensor **Z**, using a hierarchical Dirichlet process utilizing stick-breaking construction [193]. Overall, the process includes three steps which are also illustrated in Figure 4.6

1. Generate a sequence of random variables (P) from a beta distribution, given a concentration parameter α. In principle this sequence is infinite, but will be truncated to B, such that $P_b \sim \text{Beta}(\cdot|1, \alpha)$

2. Generate a probability of stick breaking (C_b) for each P_b (from step 1): $C_b = P_b \prod_{j=1}^{b-1} (1 - P_j)$

3. Generate the indicator tensor (feature and subject partitioning) from a multinomial distribution using the weights (from step 2),

A) Illustration of stick breaking

Step 1) Generate a sequence of random variables (P)

$$P_b \sim Beta(\cdot\,|1, \alpha)$$

Step 2) Generate a probability of stick breaking for each P_b

$$C_b = P_b \prod_{j=1}^{b-1} (1 - P_j)$$

$C_1 = P_1$

$C_2 = P_2(1 - P_1)$

$C_3 = P_3(1 - P_3)$

$C_4 = P_4(1 - P_4)$
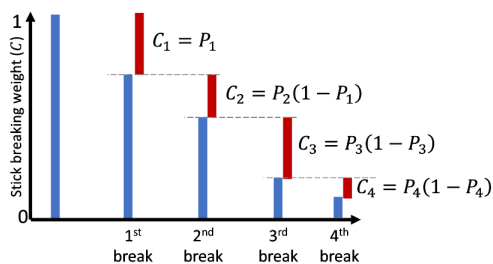
B) Notation for parameters of Dirichlet process

Notation	Description
	Dirichlet for view and feature cluster
α_1	Concentration parameter of Beta distribution (view)
w_v	Beta distributed variable for view v
π_v	Stick breaking probability for view v
α_2	Concentration parameter of Beta distribution (feature)
$w_{g,v}^{\prime(m)}$	Beta distributed variable for feature cluster g of distribution m in view v
$\pi_{g,v}^{\prime(m)}$	Stick breaking probability for feature cluster g of distribution m in view v
$\tau_{g,v}^{\prime(m)}$	Common stick breaking weight $\pi_v \pi_{g,v}^{\prime(m)}$
	Dirichlet process for subject clusters
β_1	Concentration parameter of Beta distribution (subject)
$u_{k,v}$	Beta distributed variable for subject cluster k in view v
$\eta_{k,v}$	Stick breaking probability for subject cluster k in view v

Figure 4.6: Dirichlet Process with Stick-breaking Panel A: Illustration of first two steps of the Dirichlet Process utilizing stick breaking. Panel B includes a list of the notations that are used for Dirichlet process in the remaining section. We have chosen to keep the same notation as in the earlier publications of the method, e.g. the probability of stick breaking (C in Panel A) has the notation π for the feature clusters and η for the subject clusters. Furthermore, V now refers to the number of views (and not voxels as in earlier sections)

For the feature partitioning the Dirichlet process is performed in a hierarchical structure, where views are generated first, followed by the generation of feature clusters. In contrast, subjects are partitioned into subject clusters for each view. Here it should be noted that a feature j can only belong to one of the views while subject i belongs to all views.

Feature partition tensor Y: For each distribution family m, a three dimensional feature-partition tensor $Y^{(m)} \in \mathbf{R}^{d^{(m)} \times V \times G^{(m)}}$ is modelled, where V is the number of views and $G^{(m)}$ is the maximal number of feature cluster for any view, and $d^{(m)}$ is the number of features for distribution family m . Here $Y_{j,v,g}^{(m)} = 1$ if the feature j of distribution family m belongs to the feature cluster g in the view v, and otherwise $Y_{j,v,g}^{(m)} = 0$. Combining the different distribution families gives: $Y = \{Y^{(m)}\}_m$.

The view and feature cluster membership of a feature j of family m is generated using a hierarchical Dirichlet process with stick-breaking construction to generate the variables as follows:

$$w_v \sim \text{Beta}(\cdot|1,\alpha_1), v = 1,\ldots \qquad\qquad\qquad \text{step 1 (view)}$$

$$\pi_v = w_v \prod_{t=1}^{v-1}(1-w_t) \qquad\qquad\qquad\qquad \text{step 2 (view)}$$

$$w_{g,v}^{\prime(m)} \sim \text{Beta}(\cdot|1,\alpha_2), g = 1,\ldots,G^{(m)}, m = 1,\ldots,M \quad \text{step 1 (feature cluster)}$$

$$\pi_{g,v}^{\prime(m)} = w_{g,v}^{\prime(m)} \prod_{t=1}^{g-1}(1-w_{t,v}^{\prime(m)}) \qquad\qquad\qquad \text{step 2 (feature cluster)}$$

$$\tau_{g,v}^{(m)} = \pi_v \pi_{g,v}^{\prime(m)} \qquad\qquad\qquad \text{combined probability (view and feature)}$$

$$Y_{j\ldots}^{(m)} \sim \text{Mul}(\cdot|\tau^{(m)}) \qquad\qquad\qquad \text{step 3.}$$

$$\tag{4.10}$$

Subject-partition tensor Z: Since **Z** is common to all distribution families, this implies that the model estimates the subject cluster solutions using information from all distribution families. Here $Z \in \mathbf{R}^{I \times V \times K}$ where K is the maximum number of subject clusters for all views, and I is the number of participants. $Z_{i,v,k} = 1$ if a subject i belongs to the subject cluster k and 0 otherwise. The subject cluster membership for a subject i in view v ($Z_{i,v,}$) is generated by the procedure below

$$u_{k,v} \sim \text{Beta}(\cdot|1,\beta), v = 1,\ldots,V, k = 1,\ldots,K \quad \text{step 1 (subject cluster)}$$

$$\eta_{k,v} = u_{k,v} \prod_{t=1}^{k-1}(1-u_{t,v}) \qquad\qquad\qquad \text{step 2 (subject cluster)} \qquad\qquad (4.11)$$

$$Z_{i,v,} \sim \text{Mul}(\cdot|\eta_v) \qquad\qquad\qquad \text{step 3 (subject cluster).}$$

Likelihood and prior distribution

The MCC algorithm assumes that each instance $\mathbf{X}^{(m)} \in \mathbb{R}^{I \times J}$ independently follow a given distribution conditional on the subject and feature partition tensors **Y** and **Z**. The parameters of the distribution in view v, feature cluster g, and view v of family m is denoted $\theta_{v,g,k}^{(m)}$ and the collection of parameters for all cluster blocks is $\Theta = \{\theta_{v,g,k}^{(m)}\}_{v,g,k,m}$

The log-likelihood of **X** is given by

$$\log p(\mathbf{X}|\mathbf{Y},\mathbf{Z},\Theta) = \sum_{m,v,g,k,j,i} \mathbb{I}(Y_{j,v,g}^{(m)} = 1)\mathbb{I}(Z_{i,v,k} = 1)\log p(X_{i,j}^{(m)}|\theta_{v,g,k}^{(m)}). \tag{4.12}$$

Where $\mathbb{I}(\cdot)$ is an indicator function returning 1 if the the statement is true and 0 otherwise.

Due to conditional independence the joint prior distribution of the unknown variables $\phi = \{\mathbf{Y},\mathbf{Z},\mathbf{w},\mathbf{w}',\mathbf{u},\Theta\}$ is given by

$$p(\mathbf{w})p(\mathbf{w}')p(\mathbf{Y}|\mathbf{w},\mathbf{w}')p(\mathbf{u})p(\mathbf{Z}|\mathbf{u})p(\Theta). \tag{4.13}$$

Variational Inference

Since it is computationally intractable to evaluate the marginal likelihood $p(\mathbf{X}) = \int p(\mathbf{X},\phi)d\phi$, variational inference is used to approximate the marginal likelihood by maximizing the lower bound. More specifically, the MCC algorithm uses variational Bayes Expectation-Maximization

for MAP (maximum a posteriori) where the logarithm of the marginal likelihood is approximated using Jensen's inequality:

$$\log\big(p(\mathbf{X})\big) \geq \int q(\phi)\log\frac{p(\mathbf{X},\phi)}{q(\phi)}\,d\phi = \mathcal{L}(q(\phi)), \tag{4.14}$$

where $q(\phi)$ is an distribution where the parameters ϕ are determined such that the distribution approximates the true posterior . For more details including the updating equations, we refer to the earlier publications on this method [191].

Observation models

For each subject-feature cluster block of each view, the algorithm fits a univariate distribution using conjugate priors of the specified distribution of the family. E.g. if a subject-feature cluster includes features that follow a Gaussian distribution (such as purple cluster in view of Figure 4.5) where the scaled variance has an inverse-gamma distribution, then the corresponding conjugate prior is a normal-inverse gamma distribution. The observational models, including priors, log-likelihood and update equations are specified in the supplementary section 1 of Tokuda et al. from 2017 [191]. Furthermore, the relevant equations and parameters choices for the priors are specified in Paper D.

4.6 **Performance measures**

Throughout the studies we applied different performance measures depending on the question of interest. These will shortly be summarized here, and described in more detail in the respective Papers.

4.6.1 **Study 1-3, performance measure of predictive modelling studies**

For a binary classification, the model performance is usually assessed using a summary measure based on indices of the confusion matrix which give the number of true positive (TP), true negatives (TN), false positives (FP) and false negatives (FN). For studies with balanced datasets, the accuracy (proportion of correct prediction) is a standard measure, which we have used for our task-paradigm classifications in Study 1.

However, for unbalanced datasets, the accuracy measure should be used with care, since the unbalance between classes that influences the chance model performance such that it is no longer 50%. For unbalanced classifications (schizotypy classification in Study 1 and schizophrenia diagnosis classification in Study 2) we have therefore used alternative measures which are more appropriate for unbalanced datasets, as specified in Paper B and Paper C respectively.

In regression-based prediction, the performance is measured by how close the predicted phenotypic measure is to the observed. When the phenotypic measure is a continuous outcome, the performance is often assessed using the mean squared error, or Pearson's correlation coefficient between the predicted and observed variable. In study 3, we used Spearman' s rank coefficient of correlations (Rho), which is a nonparametric measure of correlation utilizing ranks [194]. We

opted for this measure, because the PANSS scale is a summation of categorical subitems, and thus not a continuous measure.

4.6.2 Study 4: performance measures of subtyping clustering

In Study 4, we used two different kinds of performance measures: i) adjusted rand index to asses the stability of the clustering and ii) Pearson's χ^2 test to evaluate the diagnosis association of each view.

Adjusted rand index

To measure the similarity between two cluster solutions we used the adjusted rand index (ARI), which exists in the range between 0 and 1, where 1 corresponds to identical clustering and 0 implies random labelling [195]. The original rand index measures the proportion of observations that are clustered the same way in the two solutions, such that

$$\text{rand index} = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} \tag{4.15}$$

Where a and b are the number of observations where the clustering solutions (runs) agree, whereas c and d are the observations that are clustered differently as shown in Figure 4.7.
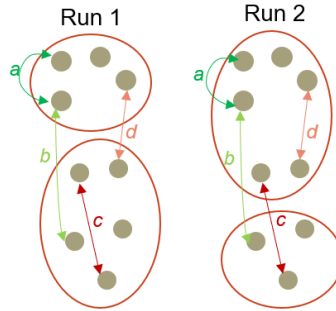
Figure 4.7: Illustration of rand index. Example of the similarity between clustering solutions (Run1 and Run2, with same data but different initialization). Here both runs result in a two-cluster solution, but they differ by their allocation of two datapoints. The rand index is then calculated according to eq. 4.15, where a: same cluster in run1 and same cluster in run2, b: different cluster in run1 and different cluster in run2, c: same cluster in run 1 and different cluster in run2, d: different cluster in run1 and same cluster in run2.

The ARI is a corrected-for-chance version of the random index, which establishes a baseline using the expected similarity of all pair-wise comparisons:

$$\text{ARI} = \frac{\sum_{r,c}\binom{n_{r,c}}{2} - \left[\sum_r\binom{a_r}{2}\sum_c\binom{b_c}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_r\binom{a_r}{2}\sum_c\binom{b_c}{2}\right] - \left[\sum_r\binom{a_r}{2}\sum_c\binom{b_c}{2}\right]/\binom{n}{2}} \tag{4.16}$$

Where $n_{r,c}, a_r, b_c$ are values from a contingency table which indexes the first clustering solution in the rows ($r = 1, \ldots, R$), and the second in the columns ($c = 1, \ldots, C$).

In Study 4, we investigated the stability of the MCC algorithm across initializations and datasplits. As a main stability measure we used the 'feature to view' assignment (ARI_{view}), but

we also used (ARI$_{subject}$) and (ARI$_{feature}$) to asses the clustering stability of the subject and feature clusterings within relevant views.

Pearson's χ^2 test for diagnosis association

To search for subtypes that were related to schizophrenia, we identified views with a significant diagnosis association by using Pearson's χ^2 test for contingency tables to evaluate the association between the subject-cluster (from clustering) and diagnosis label. The Pearson's χ^2 test statistic is used to test the independence between the row and columns of the contingency table, where independence implies that knowing the value of the row variable (here subject-cluster label outputted by the MCC clustering) does not change the probabilities of the column variable (diagnosis label), and vice versa. The Pearson's χ^2 test statistics follows an asymptotic χ^2 distribution with (R-1)(C-1) degrees of freedom, and it is calculated as

$$\chi^2_P = \sum_r \sum_c \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \tag{4.17}$$

Where $O_{r,c}$ is the observed count for the rth row and c^{th} column in the contingency table. $E_{r,c}$ is the expected counts when assuming independence (null hypothesis of the test) which is calculated as $E_{r,c} = \frac{n_{r,.}n_{.,c}}{N}$ where n_r. and $n_{.}c$ are the row and column marginal totals, and N is the total number of counts in the table.

4.6.3 Assessing significance with permutation testing

Permutation testing is a non-parametric alternative to the 'classical parametric testing'. Here the typical null hypothesis is that all observations come from the same distribution. As the variables are interchangeable under the null hypothesis, an approximate distribution of the test statistic assuming no effect can obtained by repeatedly measuring it while rearranging the observed data [196].

E.g. in a situation of binary classification, the null hypothesis is there is no difference between data from the two groups. To test this, the classification is performed both with the correct phenotypic measures (indicating the true group membership) as well as on permuted labels (e.g. using 1000 random permutations). The intuition is, that if there is no difference between the two groups, then the labelling could be arbitrary. In this way, the random permutations are used to create en empirical null distribution of the statistical measure (e.g. for classification this could be accuracy). The p-value is then found as the proportion of permuted labels which obtain a test statistical equal or larger than the one that is computed when using the correct labels.

Since permutation testing is a flexible and intuitive method which relies of minimal assumptions, it has been used for statistical inference in all our studies. When multiple test were performed, we used 'maximum permutation testing' to correct for multiple comparisons. Here we created an empirical null distribution by only considering the highest effect over the entire set (i.e. over the different comparisons). This controls the family-wise-error over the set.

In this PhD project, we used permutation testing to assess the significance of the predictions in Studies 1–3 and for the diagnosis association measure in Study 4. More details on how this was implemented is given in the corresponding papers.

RESEARCH CONTRIBUTIONS 5

In this chapter the main research contributions of this PhD project are summarized. For each study, we start with a short introduction and motivation for the study followed by a summary of the main contributions to the field. A discussion of the contributions can be found in the corresponding papers and in the next chapter. We have also added a few additional contributions that were not included in the papers, these are described in more detail.

5.1 Study 1, Prediction framework and social anhedonia

As described in section 2.5.3, the overall goal of Study 1 was to investigate if machine learning could be used to obtain significant predictions of a dataset, which had previously shown correlations between brain activation and the degree of schizotypy [148]. Furthermore, we wanted to specifically evaluate which brain characteristics would drive this prediction. To do this, we built a classification-framework that included 11 different feature extraction methods, which enabled us to investigate the separate importance of static, temporal and spatial network features for prediction. To validate the classification-framework and the utility of decomposition features, we performed a task-paradigm classification for which the exact label was known (e.g. theory of mind vs. physical control condition task). The findings from this study were published in Paper B, and furthermore we wrote a perspective paper about the topic (Paper A).

Contributions

Perspective paper on potentials and challenges: To our knowledge, Paper A was the first 'fMRI machine learning review and recommendation' publication in the schizotypy field. It includes a review of earlier fMRI studies in schizotypy (both using univariate brain mapping and machine learning classification) and discussed the potential and challenges of data-driven machine learning approaches. We also commented on best practices of procedures for future studies to provide specific recommendations on how to plan a machine learning study to predict schizotypy traits.

Expansion of MSAA to enable spotlight analysis: We implemented the spotlight approach to the MSAA model, which had previously been described by Hinrich et al. [165]. The goal of the spotlight approach is to restrict the seed region matrix to predefined regions of interest, which forces the method to find networks related to these brain regions. We hypothesized that the spotlight restriction would result in components that could obtain better classification performance. However, our results showed that the whole brain MSAA (without spotlight restrictions) obtained superior classification for both the task-paradigm and social anhedonia prediction.

Stability analysis of MSAA: As other decomposition methods, MSAA is a non-convex optimization problem, which means that the final solution can change depending on the initializa-

tion. For the whole brain MSAA model, we showed that the brain networks found by the MSAA algorithm were quite stable when repeated initializations were used. We ran a stability analysis, in which we repeated the MSAA decomposition with 100 different initializations, which we divided into 10 groups of each 10 runs. We compared the stability of the spatial networks when the best (lowest cost function of 10 runs within each group) were compared across the 10 groups. Using this procedure, we found that the mean stability (averages over networks and participants) of the networks was 86% compared to 80% when the decomposition was not repeated 10 times.

Validation of classification-framework: We validated the utility of using our classification-framework through a task-paradigm classification for which the true label was known. Here, we found that both temporal and spatial network features enabled significant classification of the ongoing social cognition task, but that this was not possible with static brain features. Interpretation of the predictive components (those that obtained significant classification) showed that they included brain regions that are important for theory of mind and empathy processing [149, 150].

Classification of social anhedonia: We found that it was possible to classify patients with high and low social anhedonia when using features from the decomposition methods (both temporal and spatial) and the parcellation based connectivity analysis. Both ICA and MSAA found very similar brain networks to have the highest performance. These networks included regions that had also previously been related to social anhedonia [148].

In summary: Using a broad range of feature extraction methods, we found that significant classifications were obtained when using temporal and spatial network features, and that the best performances were obtained with features from the decomposition methods. We showed that the novel MSAA extracted stable brain networks, which were similar to those extracted by ICA, and we successfully implemented the spotlight approach to the MSAA model.

Throughout the analyses of Study 1, we discovered how much the final results depended on the parameters within the analysis pipeline, all the way from prepossessing, to parameters of the feature extraction and classification. Thus, we considered Study 1 as an exploratory investigation of features for classification, and highlighted that it would be important to validate our findings in external data to draw conclusions on their validity [166]. This was also described by other studies that were published around that time, which showed that even when using cross validation, studies could still be overfitted, and that the limited reproducibility that was found by univariate brain mapping studies, was also a problem for predictive modelling studies [74, 176].

5.2   Data and features used in Studies $2 - 4$

For the remaining studies we thus focused our analysis on efforts to increase the robustness and reproducibility. One important step towards this was the use of multi-site fMRI data, which had been made publicly available through data sharing initiatives since my first PhD enrollment. Furthermore, we adapted our prediction framework to make it more robust, e.g. by using fMRIprep for data-derived preprocessing to obtain a more streamlined and robust processing across sites. Furthermore, we opted for focusing our further analysis on connectivity features (both in the

form of parcellation based connectivity matrices and spatial maps from decomposition methods), which consistently had proved superior to static features in Study 1. We did not continue with the investigation of temporal features, since we now used resting state data.

5.3   Study 2, Diagnosis classification of multi-site data

Many earlier studies have investigated differences in functional brain activation between patients with SZ and HCs, but few findings have been tested on external data, and for those that have, the results showed substantially lower prediction performances on the external data [15, 21, 66, 197]. In addition, even though most recent studies demonstrated that patients with schizophrenia have hypoconnectivity in a large part of the connectome, firm conclusions are still to be drawn on what brain regions and networks that drive the differences, and there are so far no clinically used biomarker to inform diagnostic decisions [6, 13, 198].

The goal of Study 2 was to use multi-site connectivity features for diagnosis classification of schizophrenia and to investigate ways to increase the robustness and reproducibility of the developed models. All models were developed and optimized on a multi-site discovery dataset (D2a) and the reproducibility of the final models were tested on an external dataset (D2b) comprising data from two sites that were not used for model training.

Based on our experiences from Study 1, we chose to expand our analysis of the **decomposition features** by considering different ways of 'transfer learning' and by creating 'ensemble decisions' across brain networks. We studied the effect of transfer learning by applying the decomposition methods (ICA and MSAA) in three different ways. In approach 1, decomposition was performed independently on D2a and D2b, in approach 2 it was performed on the combined dataset, and in approach 3 we used the decomposition output from the discovery dataset to extract features from the external test dataset. We then investigated how these different transfer learning approaches influenced the stability of the networks and classification performance. This was important since there is so far no consensus on how decomposition features should be used for predictions across datasets. Furthermore, whereas we in Study 1 had performed the classification of each individual brain network, we now also implemented ensemble models, which made one common decision based on the predictions of the individual networks.

To increase the interpretability of the **parcellation based connectivity features**, we performed a post-hoc analysis with repeated classifications on 'subparts' of the connectivity matrix (only including features that were related to individual RSNs), to investigate if any of the RSN could drive the classifications by itself. To our knowledge using separate 'subpart' classifications, have not previously been performed, and we compared the results of these classifications with the contributions as estimated by classification weightmaps.

Contributions:

MSAA on multi-site rsfMRI data: We showed that MSAA could be used to extract stable RSN, which were similar to those found by ICA, even when it was applied on multi-site data. Due to the large number of participants (and features), the time-complexity of the MSAA was high

($\mathcal{O}(VTKI)$, V: voxels, T: time points, K: number of components, I: number of participants, as described in Hinrich et al. [165]). A way to reduce the computational and time complexity is to perform PCA prior to the decomposition analysis, similarly to what is typically done in ICA. We therefore repeated the MSAA analysis where we performed a PCA analysis prior to MSAA (applied in the temporal dimension for which the original data included 170–260 time frames (depending on the site)). In this analysis we included 35 components for further analysis, since this was the same number as we used in ICA as estimated by the minimum description length [159], which was confirmed by visually inspecting a plot of the eigenvalues . By that we reduced the computational complexity with a factor of five, and the time complexity with a factor 2, the latter was not improved with a factor of five since the 'PCA MSAA' model needed approximately twice the number of iterations before it converged. Overall, MSAA found similar RSNs with and without the preceding PCA step, as illustrated in Figure 5.1. For the remaining part of the study we continued to work with the full MSAA (no PCA) since this was consistent with the version we previously used and performed stability analysis for. For future studies, we believe that including a PCA step prior to MSAA can be a valuable tool to enable more efficient inference. It might even be necessary step in cases where the available memory is insufficient to perform optimization on all available data.
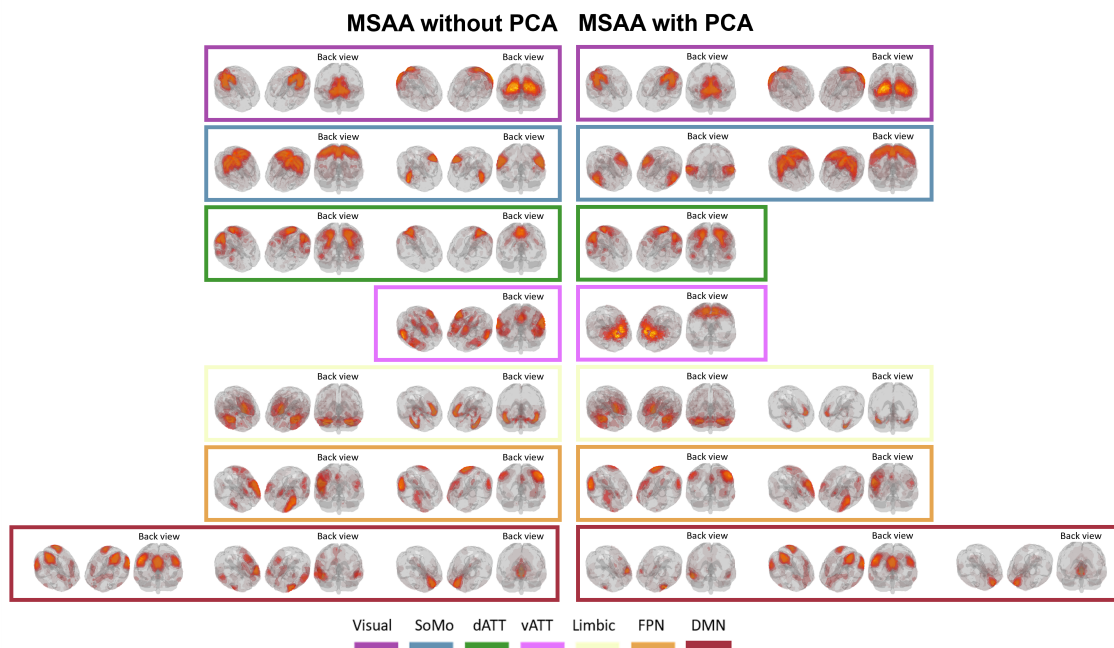
Figure 5.1: Comparison of MSAA networks with and without PCA. Spatial resting state networks determined by MSAA with (right) and without (left) a PCA preceding step. The networks are sorting according to the the 7-RSN parcellation given by the Yeo et al. [118], as described in section 3.7. The number of iterations were 129 for the MSAA model without PCA and 226 iterations for the model with PCA. 3D illustration produced using the VITLAM toolbox which is available via GitHub at (https://github.com/JesperLH/VITLAM) [165].

Transfer learning for decomposition networks: Comparing the different transfer learning approaches for the decomposition methods, we found that transfer learning approach 3

was superior for both MSAA and ICA. With this approach, we found that the networks were stable across datasets, obtained superior classification performance, reduced the computational complexity and ensured a direct matching of components between datasets.

Reproducibility on external data for our ensemble models: For all three feature extraction methods we found that the ensemble models generalized to the external dataset. Comparing the performance of the features from the parcellation based connectivity analysis and the ensemble prediction of the decomposition methods, we found that they were similar on the discovery dataset, but that the parcellation based features obtained higher performance on the external test data, possibly because they included both within and between RSN information.

Importance of individual RSNs: Overall, we found that there was no 'single-best' RSN classifiers (with substantially higher performance than the remaining networks), and that the best performance was obtained when using ensemble models that included data from all networks. For the decomposition methods, the importance of each RSN was investigated by performing the classification on individual spatial maps. Here, we found that it was not the same networks that obtained best classification performance across the two datasets. We found, however, that RSNs within the somatomotor, visual and ventral attention networks were consistently among the best performing networks for both ICA and MSAA.

For the parcellation based connectivity analysis, we investigated the importance of different RSNs using weightmaps and 'subpart' predictions. Figure 5.2 shows the results of within and between RSN connectivity for each of the seven RSNs when using the weightmaps (top) and subpart classifications (bottom). Overall, we found the same pattern for the two approaches, where the *between* RSN connectivity shows a higher weight/classification performance, compared to *within* RSN contributions. Both methods found the highest performance for the limbic RSN; however, not with a substantial margin compared to the other RSNs. As for earlier studies, we found that patients with SZ had hypoconnectivity (i.e., reduced connectivity between regions) within the RSNs, whereas the between RSN connectivity was more mixed and included both hyper- and hypoconnectivity (shown in Supplementary Figure 4 of Paper C) [13, 88].

In summary: In Study 2 we were able to build a classification model on multi-site rsfMRI data, which generalized to the external dataset. To our knowledge, this is one of the first studies to show a high and reproducible classification performance across datasets. Comparing the different transfer learning approaches we found that approach 3 was superior on all parameters, and we thus suggest this approach for future studies that want to use decomposition methods across different datasets. The highest classification performance was found when using ensemble prediction models, which supports earlier findings that schizophrenia affects a wide range of brain networks.

5.4 Study 3, Prediction of PANSS scores on multi-site data

The overall goal of this study was to determine if the features from Study 2 could also be used to predict the symptom severity (PANSS$_{total}$) and three PANSS subscale, in an attempt to tackle the internal heterogeneity of schizophrenia [21]. For the decomposition method, we only used

A) Absolute Weightmap

Visual SoMo dATT vATT Limbic FPN DMN

B) Mean weight (z-scored)

	Visual	SenMo	dATT	vATT	Limbic	FPN	DMN
■ B	4.6	4.5	4.4	4.0	4.6	4.8	4.6
■ W	4.1	3.9	4.0	4.2	4.1	4.0	3.9

C) Average FC

D) Classification performance for each network

	Visual	SenMo	dATT	vATT	Limbic	FPN	DMN
■ B	0.84	0.85	0.84	0.86	0.90	0.84	0.85
■ W	0.64	0.67	0.60	0.68	0.82	0.66	0.70
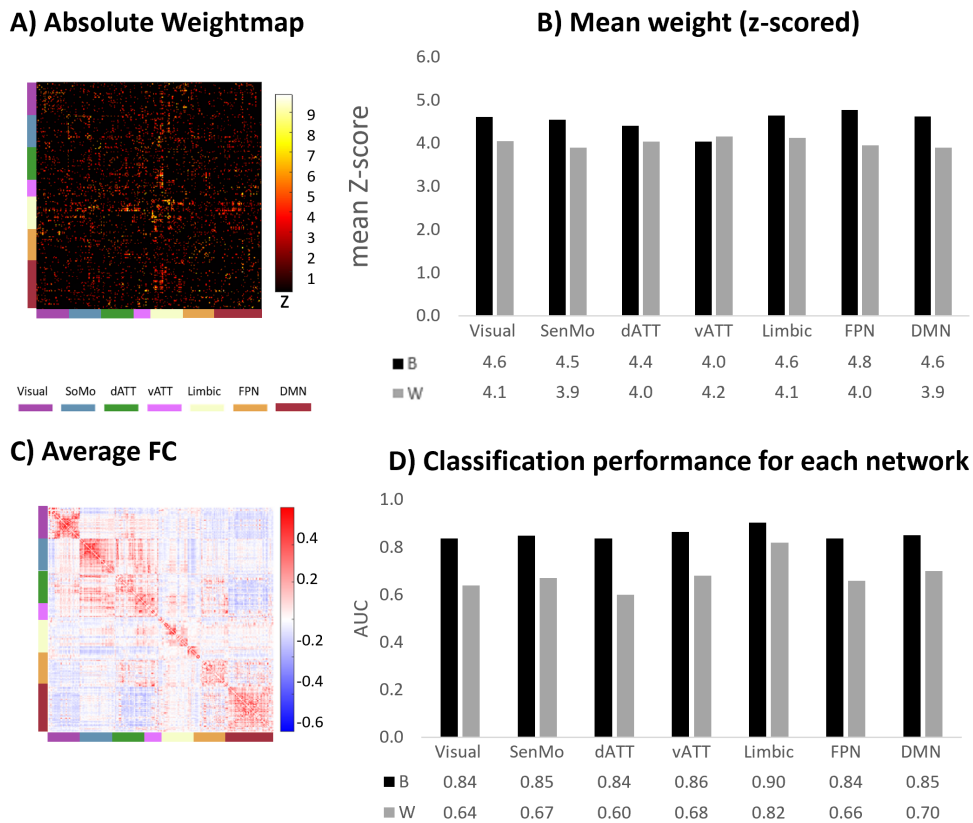
Figure 5.2: Comparison of RSN contribution with weightmaps and subpart classification.
The top panel shows the results of using the weightmap (adjusted version of Supplementary
Figure 3 of Paper C. The weightmap was adjusted using the data covariance as suggested by Haufe
et al. [199]. The methods are described in supplementary material of Paper B). Panel A shows the
absolute weightmap and Panel B shows the mean weight for each individual RSN, where black
and grey bars indicate between and within RSN connectivity contributions, respectively. The
bottom panel shows the results from subpart classification analysis (this is equivalent to Figure 5
in Paper C) where Panel C shows the average connectivity matrix (over participants) and Panel D
shows the classification performance of each RSN.

RSNs from transfer learning approach 3, and we also opted for using Gaussian process regression
(GPR) instead of support vector machines (as in the first two studies), since GPR require less
hyperparameter tuning and provide certainty estimates for the predictions. For the decomposition
methods, we used these certainty estimates to make an ensemble decision model across the RSNs,
as illustrated in Figure 2 of Paper C.

To our knowledge, there are only few earlier studies that used machine learning to predict the
PANSS scores of fMRI data, and none of these have used decomposition methods nor multi-site
data [119–122, 200]. Several of the earlier studies showed that the between RSN connectivity is
more important for the PANSS prediction than the within RSN connectivity [120, 121]. Initially
we therefore planned to study within and between RSNs contributions as we had done in Study
2; however, since the PANSS predictions were low to moderate, and did not generalize to the
external data, we did not perform any post-hoc analysis.

Contributions

Predicting symptom severity: For the total PANSS scores all three ensemble methods found a moderate prediction on the discovery dataset. However, the predicted values only resembled a positive trend around the mean PANSS score, meaning that patients with high observed PANSS score tended to have slightly higher predicted PANSS scores, and vice versa. The difference between high and low predicted PANSS scores were, nevertheless, small. Additionally, we found that these 'trend-like predictions' generally did not reproduce to the external dataset. For the individual RSNs we found a similar pattern but with even lower prediction performances. The best prediction was obtained when using the ventral attention (vATT) network. Interestingly this was found for both the ICA and MSAA analysis.

PANSS subscale prediction: As for the $PANSS_{total}$ the predictions of the PANSS subscale had moderate performance on the discovery dataset and generally did not reproduce on external data. For the prediction on individual RSNs we found that it was not the RSNs with the highest prediction performance on the discovery dataset that performed the best on the external data. This was particularly clear from the negative PANSS subscale. This finding relates to the "multiple comparison paradox" described by Marek et al. [71], who found that correcting for multiple comparisons (and thereby choosing the solution with the highest performance) reduced the probability of successfully replicating results.

In summary: We found that the prediction performances were low to moderate on the discovery dataset, where the predictions resembled a positive trend around the mean PANSS score. Furthermore, the predictions did not generalize to the external data. This was the case both for the ensemble prediction models and individual RSNs for the decomposition methods. The moderate prediction performance and poor generalizability indicate that the study might have been underpowered [71] or that differences between sites were too large compared to the signal of interest. Other potential explanations could be the internal consistency of the PANSS itself, or that the applied method or even datatype (resting state connectivity) might not be the right path forward to find robust biomarkers as discussed in Paper C.

5.5 Harmonization of multi-site data (Related to Study 2 and 3)

In Study 2 and 3 we kept the prediction analyses as data-driven and robust as possible, e.g. in the preprocessing we used a simple and robust pipeline that did not make any adjustments with respect to site differences. The motivation for this was to search for biomarkers that did not need site specific adjustments, but which could directly generalize to external data. However, earlier studies have shown that post-acquisition multi-site harmonization of brain features can be an advantage, and potentially even needed since the site-specific biases (as described in section 2.4.1) can be even larger than disease related factors [77]. This section includes results from a preliminary multi-site harmonization analysis that we performed using the ComBat method as described in section 3.9. As our initial investigations indicated that the improvement of applying

harmonization was limited, we opted against performing a principle analysis.

We applied ComBat harmonization on the entire discovery dataset, as illustrated in Figure 5.3. The inputs to the ComBat algorithm were the connectivity data, a site-indicator that specified the site of each participant and a design matrix that included the known effects of interest (sources of variability) that should not be 'removed' during harmonization.
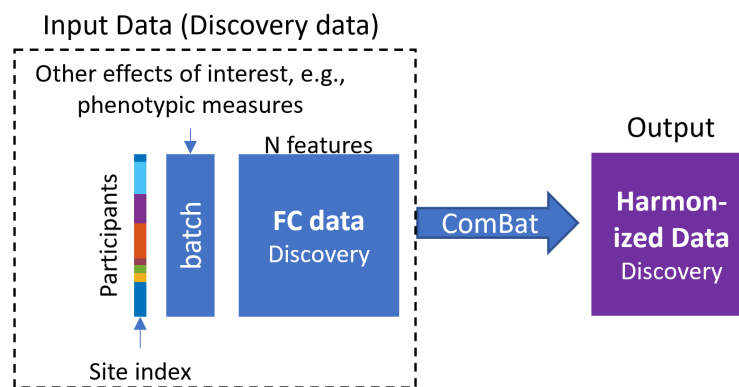
Figure 5.3: Framework for ComBat harmonization. The inputs to the ComBat algorithm were the original data, a site indicator that specified the site of each subject and a design matrix which included the known effects of interest that should not be removed during harmonization. The output was a harmonized dataset with the same size as the input data.

5.5.1 Effect of multi-site harmonization on the brain features

We applied the ComBat harmonization to the connectivity matrix from the parcellation based analysis. Figure 5.4 illustrates the effect of the ComBat harmonization on the data. Panel A shows the direct effect of the harmonization on the data, where the connectivity matrix is concatenated (rows) for each participant (columns). The colorbar on the top indicates the site allocation of each subject. To the right, the difference map between the original data (left) and ComBat harmonized data (middle) is shown. Here, it is evident that most of the connectivity values were not substantially affected by the harmonization procedure (difference is close to 0), while the maximal harmonization values reached up to approximately $Z = \pm 1.5$. To increase our understanding of how the connectivity between regions was affected by the harmonization, Panel B in Figure 5.4 shows the average (mean over participants) connectivity matrix. Here, the average harmonization is close to zero for each connectivity feature (max $Z = \pm 0.02$), indicating that none of the features were systematically affected by the harmonization procedure.

Another way to measure the effect of the harmonization is to assess the effect size (here measured using Cohen's D) of the site effect as described in section 3.9. Figure 5.5 shows the effect size before (left) and after (right) ComBat harmonization, for each of the eight sites that were included in the discovery dataset. Here we found that the mean Cohen's D was already low (below 0.4) on the original data, and that the ComBat harmonization mostly reduced site effect of 'outliers in the connectivity values' that displayed a large site effect on the original data.
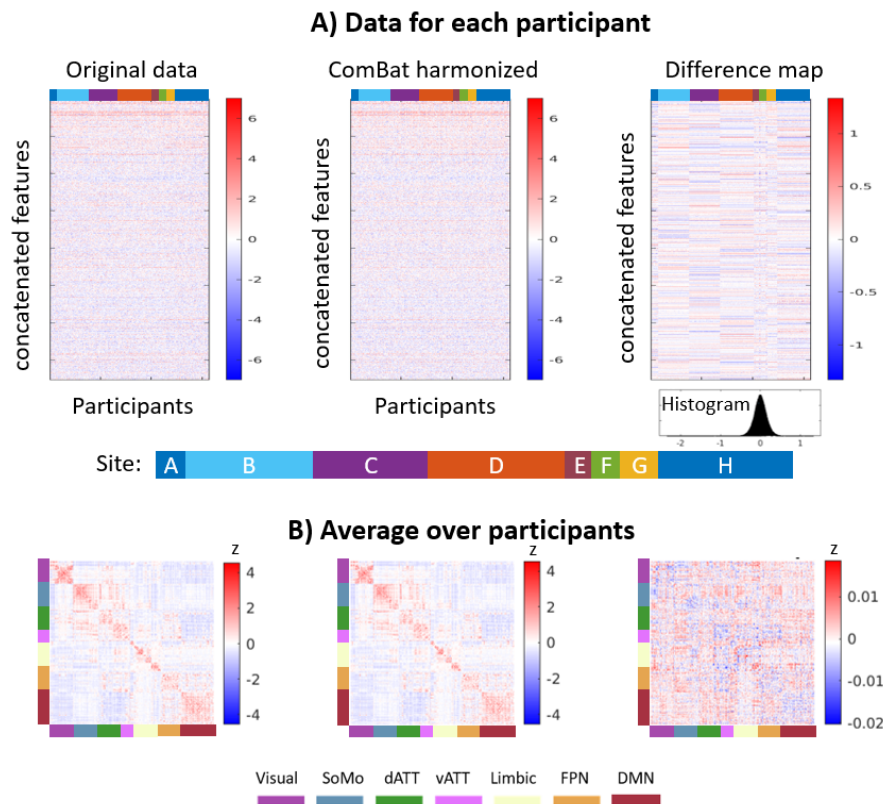
Figure 5.4: ComBat harmonization on connectivity matrix. Panel A illustrates of the data (features as columns and participants as rows (colorbar on top indicates the site index of each participant), both before(left), after ComBat harmonization (middle) and a difference map (Original-ComBat) (right). Below the difference map is a histogram that shows the distribution of the differences. Panel B illustrates the mean (over participants) connectivity matrix arranged by the seven RSNs (original left, ComBat middle and difference to the right).
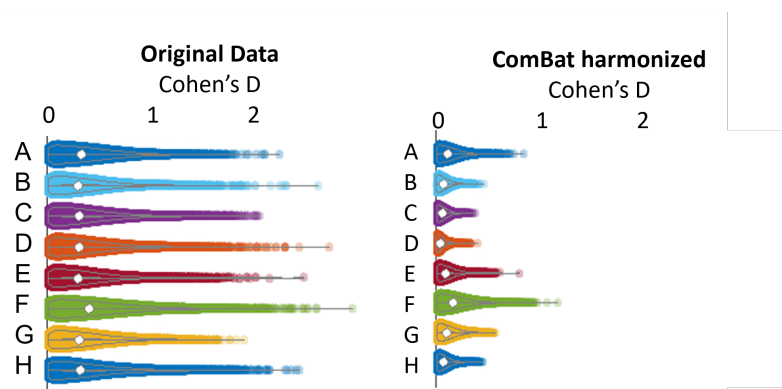
Figure 5.5: Cohen's D of site effect before and after ComBat harmonization. Violinplot of the effect size (Cohen's D) for each of the eight sites (A-H) that were included in the discovery dataset. The original data (no ComBat) is shown to the left, and the right panel shows the effect size after ComBat harmonization.

5.5.2   **Effect of multi-site harmonization on the prediction results**

The ultimate goal of multi-site harmonization is not only to reduce the site effect, but to increase prediction performance on external datasets by removing site specific measurement biases on the training (discovery) dataset. The rationale here is that when a model is trained on data that includes less measurement bias, then it can better pick up signals that are related to the phenotypic measures of interest, such that it can also predict these more efficiently when applied on another (external) dataset.

As illustrated in Figure 5.3 the ComBat harmonization was applied on the whole discovery dataset at once. This implies, that when the prediction performance on the discovery dataset is measured using cross validation, this estimate will be biased, as the harmonized data was informed about the phenotypic measure of interest (e.g. diagnostic label) as part of the harmonization. This is a clear example of when the training and test datasets are no longer independent and include factors that will bias the performance of the prediction. This could have been mitigated by applying the harmonization within each cross validation loop. However, since our goal was to determine how the ComBat harmonization on the discovery dataset influenced the prediction performance on the external dataset, we opted against this here to reduce the computational complexity.

Figure 5.6 shows the prediction performances on the discovery dataset (top) and external dataset (bottom). We found that the ComBat harmonization on the discovery data did not increase the performance on the external data, neither for the classification nor the $PANSS_{total}$ prediction analysis. Furthermore, even for the discovery dataset (where the prediction on the harmonized dataset was biased as described above), the prediction performance was only increased for the diagnostic classification (left panel), whereas the performance on the $PANSS_{total}$ prediction was actually reduced after harmonization.

The results presented here are valid for the setting where supervised machine learning is used for predictive modelling. Arguably in this setting the algorithm actually automatically seeks to disregard any site differences given that it is trained on multi-site data. The conclusion is likely to be different in cases where a parametric model is used to test for statistically significant differences between sites.

In summary: Based on these somewhat preliminary results, we opted against dedicating further efforts to investigate harmonization procedures for the remainig analyses. However, we believe that more structured investigations of harmonization methods would be beneficial (e.g. also including methods that rely on data from travelling subjects) to further investigate the potentials and limitations of these methods.

5.6   **Study 4, Subtyping using multiple co-clustering**

Whereas Studies 1–3 focused on supervised machine learning to predict phenotypic measures, the goal of Study 4 was to use unsupervised clustering for disease subtyping. In Study 3, we made an attempt to address the internal heterogeneity by predicting the PANSS subscale scores instead of classifying the diagnostic labels, but we did not identify any clinically meaningful predictions.
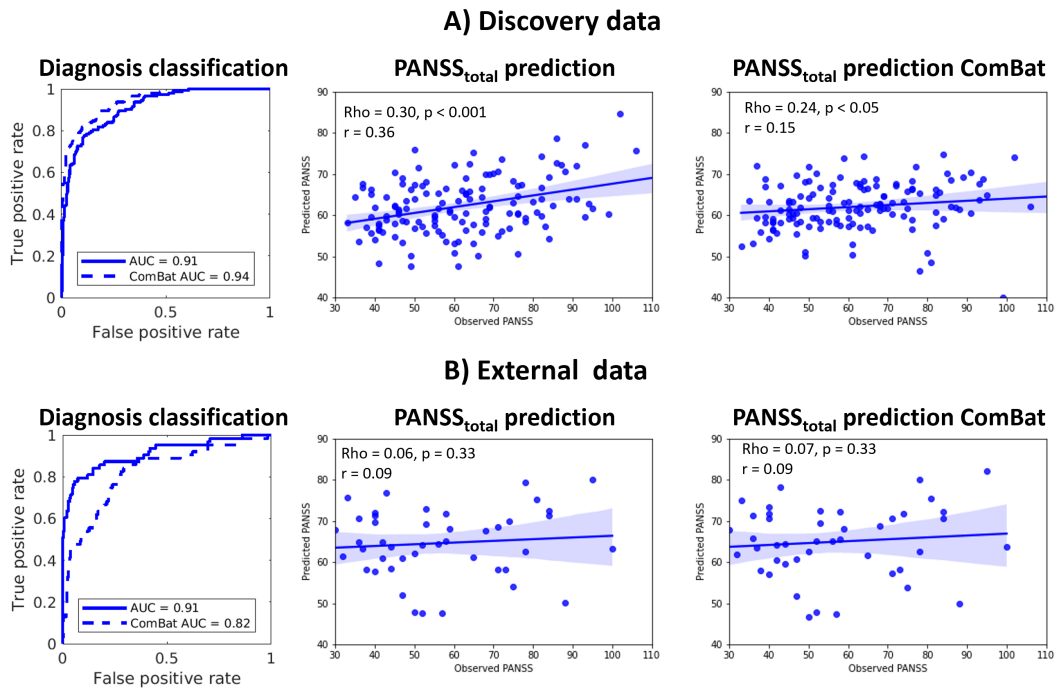
Figure 5.6: Prediction results before and after ComBat harmonization. Prediction performance on the discovery dataset (top) and external dataset (bottom), for the diagnosis classification (left), and PANSS$_{total}$ prediction (middle and right). The panel in the middle shows the results of the PANSS$_{total}$ prediction without harmonization (same figures as in Paper C).The scatter plots show the predicted PANSS (y-axis) as and the observed PANSS (x-axis), where the line shows the linear regression, and the shaded area indicates the standard error of the mean. The primary prediction performance was Spearman's rank coefficient of correlation (Rho) and we also listed the Pearson's correlation coefficient (r) as this is often used in earlier studies.

One possible explanation (which is discussed in Paper C) is that the PANSS scores did not capture the main sources of the heterogeneity within the data, which highlights the importance of more data-driven approaches to identify subtypes with a more homogeneous biology.

The performance of prediction models are typically evaluated by how close the predicted outcome is to the observed value; however, measuring the performance of clustering is more centered on the ability to separate the data and the stability clusters. If a clustering solution is stable across different initializations and changes in the datasets, it is more likely to reflect a meaningful structure in the data, than if the solution is variable (unstable) across these factors. A large part of our analyses were thus focused on determining the stability across initializations and changes in the datasets. To our knowledge, this was the first time that a study systematically compared and reported how these factors affect the stability in a realistic setting on real fMRI data. We used the adjusted rand index (ARI) to measure the similarity between runs, with the 'feature to view' assignment (ARI$_{view}$) as our primary stability measure, but we also used (ARI$_{subject}$) and (ARI$_{feature}$) to asses the clustering stability within relevant views.

We used the multiple co-clustering (MCC) algorithm [191] since it has attractive properties for fMRI data (e.g. multi-view and co-clustering abilities, polytopic learning and being able to

handle missing data) and because it previously showed promising results for subtyping patients with major depressive disorder [192]. We applied the clustering on a balanced dataset, with an equal number of patients with SZ and healthy controls. The inclusion of healthy controls gave us the possibility to select subtypes based on their 'diagnosis association', since it is otherwise not trivial to assess whether specific characteristics that differentiate samples are indeed associated with the disease rather than to subgroups of the general population. In principle the clustering algorithm could be applied on all the spatial network features that we have investigated, but we here opted for using connectivity features from a parcellation based analysis, where we used the RSN parcellation presented by Allen et al. [153]. Furthermore, we performed a preliminary analysis of how the stability was influenced by using a larger atlas (more brain regions and thereby connectivity features). In this analysis, we compared the stability of the Allen atlas (results in Paper D) and on the 275ROI atlas, which was also used in Studies 2–3.

Contributions

Stability across initialization: When using connectivity features from the Allen atlas, we found that the view and feature clustering stability was high, while the subject clustering stability was only moderate. Furthermore, we found that there were still occasional abrupt changes in the cost function (log-likelihood over iterations) close to the final number of iterations (1000 iterations) indicating that the solution had not yet fully converged. Since earlier publications have not reported the number of iterations (default number chosen by the method is 30) nor the cost functions, we could not compare this finding to earlier studies. In future studies, we strongly recommend to run additional investigations of how the stability of the clustering depends on the number of iterations.

   Stability across data splits: As expected, we found that the stability across data splits was substantially lower than across initialization, which indicates that the solution is highly variable depending on variations in the set of observations (participants) that are included.

   Stability across atlases: Given the slow convergence of the algorithm and our results from the initial stability analysis, we expected that the stability of the algorithm would also depend on the size of the input data, i.e. the number of connectivity features included for each participant. To test this, we conducted stability analysis 1 (initialization) and 2 (data splits) for an additional atlases with more ROIs.

   Figure 5.7 shows the setup for this analysis. First two datasets were created using either the Allen RSN atlas [153] (28 ROIs = 378 features) or the 275ROI atlas (same as in Studies 2-3 with 275 ROIs = 37.675 features) [151]. We hypothesized that the clustering would be more stable on the atlas with less ROIs (Allan atlas).

   Figure 5.8 shows the stability results for the 275ROI atlas, while the results for the Allen atlas are shown Figure 3 and 4 of Paper D. As expected we found the MCC was more unstable on the high dimensional 275ROI atlas, where the mean ARI_{view} dropped from 0.84 (Allen atlas) to 0.48 (275ROI atlas). We could not evaluate the subject and feature cluster stability for the 275ROI atlas, since it was not possible to meaningfully match views between runs as the the feature to view assignment was too low. This clearly shows that the clustering solutions was very unstable even
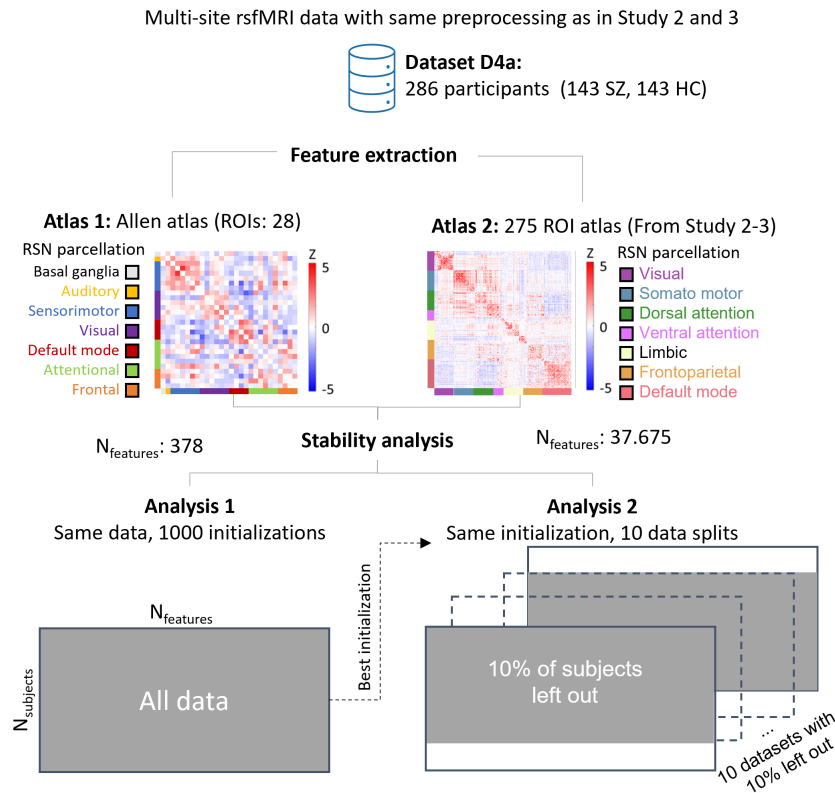
Figure 5.7: Subtyping stability analysis over two atlases. Parcellation based connectivity features were extracted using two different atlases, 1) The Allen atlas which included 28 ROIs [153]and 2) the 275ROI atlas which was the same as we used in Study 2 and 3 [151]. For features of both atlases we performed stability analysis 1 and 2, to determine how the size of the input feature space affected the stability across initializations and data splits.

across initializations when using such a high dimensional atlas. For stability analysis 2 (data splits), we found that the mean ARI_{view} was similar (0.48 (Allen) and 0.42 (275ROI)) across the two datasets, indicating that the clustering is relatively unstable across data splits for both atlases. The remaining analyses were therefore performed using features from the Allen atlas.

Views with disease related subtypes: For the best solution (assessed by the stability and log likelihood) of the Allen atlas, we found one view with a significant association between the diagnostic label and the subject clustering solution as illustrated in Figure 5.9.

That view included eight subject clusters, where each cluster (potential subtype) had a varying proportions of patients with SZ (from $15 - 73\%$). Three of these had a predominance of patients with schizophrenia (subject clusters $6-9$) and could be described a "schizophrenia related subtypes". However, it should be noted that there were both healthy controls and patients with schizophrenia in all subject clusters. Furthermore, the view included three feature clusters, where feature cluster 2 and 3 showed a linear trend in their connectivity values as illustrated in Panel A of Figure 5.9. I.e., feature cluster 2 mostly included negative connectivity values for subject clusters that had a predominance of healthy controls, and positive connectivity values for schizophrenia-related subject clusters, and vice versa for feature cluster 3. With respect to brain
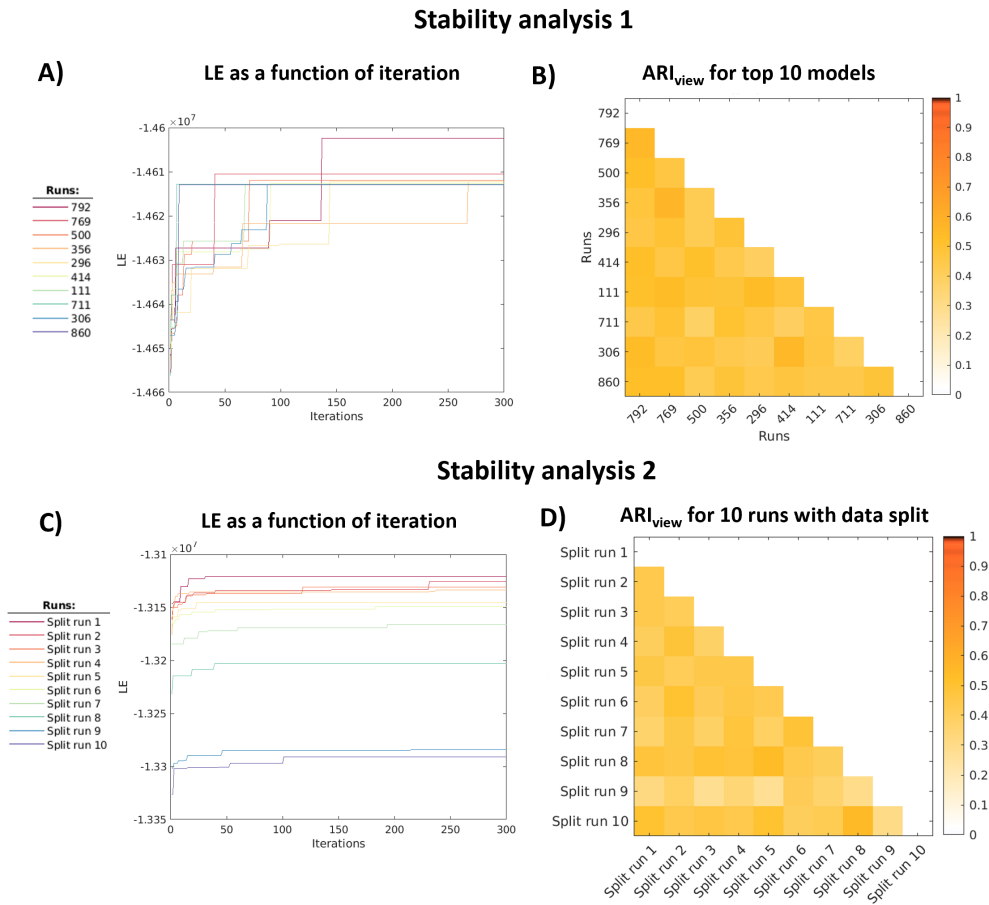
Figure 5.8: Stability analysis of 275ROI atlas. Results from stability analysis 1 (top) and 2 (bottom) when using the 275ROI atlas. For stability analysis 1, the mean ARI_{view} = 0.48, and ARI_{view}= 0.42 for stability analysis 2.

regions of the different feature clusters, we found that all six connectivity features included in feature cluster 3 were related to the basal ganglia RSN (illustrated in Panel B of Figure 5.9). While feature cluster 1 and 2 included more mixed RSN, however still, many connectives that were related to the basal ganglia RSN (Figure 6 of paper D). The basal ganglia RSN of the Allen atlas included several subcortical regions, including the striatum, which earlier has been described to have a core role in schizophrenia, particularly in relation to the dopamine hypothesis and positive symptoms. However, since the Allen atlas did not provide a more fine-grained representation of regions within the basal ganglia RSN, we can not make firm conclusions on the contribution of the striatum (compared to other brain regions in the basal ganglia). More details on the clustering solution are described in Paper D.

Subject-cluster separability: For the view with a significant diagnosis association, we performed a subject-cluster separability analysis, focusing on the separability of the clusters for each of the included feature clusters (instead of averaging them as in earlier publications). We found this to give valuable additional information that was important for the interpretation of the feature clusters, since the subject clustering differences were not the same across feature
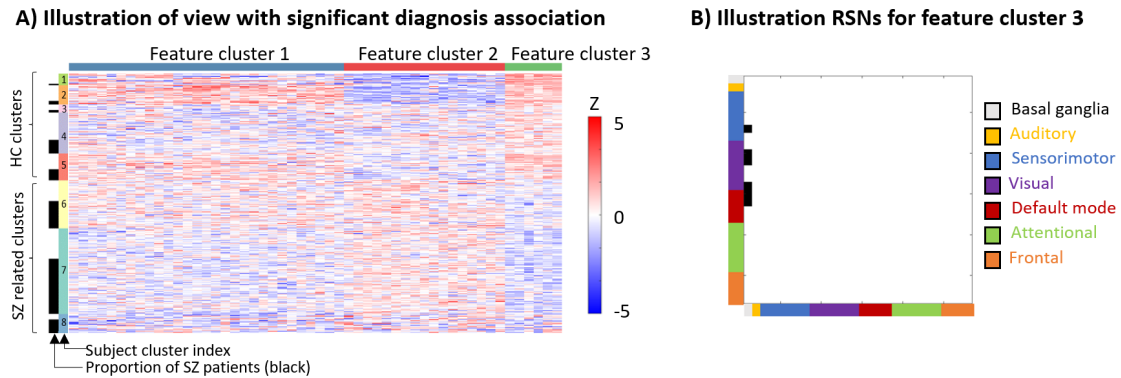
Figure 5.9: Illustration of the view with significant diagnosis association. Panel A illustrates the view with the significant diagnosis association, which includes eight subject clusters and thee feature clusters as indexed with the vertical and horizontal colorbars respectively. Subject clusters 6−8 are referred to as schizophrenia (SZ) related clusters, since they include a predominance (>50%) of patients with schizophrenia. Furthermore, these clusters also included all patients with a PANSS score above 75 (moderately affected by schziophrenia [201]), and the subject cluster separability was high between cluster 5 and 6, particularly for feature cluster 3. This feature cluster showed a continuous-trend with positive connectivity values for HC clusters, and negative connectivity values for SZ related clusters. Panel B illustrates that all RSN connectivity features included in feature cluster 3 included the basal ganglia RSN, which includes several subcortical regions, including the striatum.

clusters. We found high subject cluster separability between subject-cluster 5 and 6, which was particularly strong for feature cluster 3. This result supported the same schizophrenia-related clustering as we found by the proportion of patients with SZ as described above.

Correlation to PANSS scores: To evaluate if any of the data driven feature clusters were related to clinical factors assessed by the PANSS scale, we performed a correlation analysis. First of all, we found that patients within a subject clusters were grouped horizontally (i.e., similar mean functional connectivity) which is accordance with our expectation since subject clusters are formed based on participants with similar feature distributions. Furthermore, we found that all patients with a PANSS$_{total}$ above 75 (considered "moderately affected by schziophrenia" [201]) belonged schizophrenia-related clusters. However, it should be noted that these clusters also included healthy controls and patients with a lower PANSS score. Lastly, we found that none of the feature clusters were reproducibly correlated with any of the PANSS scales. This shows that the feature clusters did not directly relate to any of the clinical representations that were available through the PANSS scale, and indicates that the clustering solutions reflect other sources of variability in the data.

Reproducibility of diagnosis association in subtypes: In our validation analysis we showed that the diagnosis association of the subject clustering described above reproduced on the external dataset. To our knowledge this is the first study that has shown a data driven subtyping on schizophrenia patients that generalized across datasets. Furthermore, we found that the continuous trend in the connectivity values were also reproduced on feature cluster 3, and we again found support for the three schizophrenia-related subject clusters.

In summary: We found that the clustering solution of the MCC model was highly dependent

on the changes in the fMRI connectivity dataset, and that the clustering became quite unstable when an atlas with too many features were used. For the Allen atlas, we found that the feature to view stability was high, but that the subject and feature clustering solutions were only moderately stable across initializations. Nevertheless, we were able to find a subtyping solution which had a significant diagnosis association both on the discovery and external dataset. We found three schizophrenia-related clusters, for which the subject clustering had a predominance of schizophrenia patietns on both datasets. Furthermore feature cluster 3, for which all connectivity values were related to the basal ganglia RSN, showed a continuous trend on both datasets. Finally, none of the feature clusters were reliably correlated to any of the PANSS scales, which indicate that these reflect other sources of variability in the data.

We see these findings as very promising steps, and consider subtyping methods, such as MCC, to have a great potential towards exploring more data-driven disease subtypes with a more homogeneous biology. For example, it would be very interesting to further analyse the potential schziophrenia related subtype solution found in feature cluster 3, e.g. by expanding the atlas to include more subcortical regions to investigate the role of the striatum compared to other regions in the basal ganglia.

DISCUSSION AND FUTURE DIRECTIONS

6

Throughout the studies of this PhD project we have explored different ways to use machine learning and (multi-site) fMRI imaging for robust feature extraction, predictive-modelling and disease subtyping. In this chapter, we highlight some of our contributions, which we see as important methodological contributions towards discovering reproducible fMRI biomarkers for schizophrenia, and share our thoughts on future directions.

Feature extraction

Even though the choice of feature extraction heavily influences the outcome of the study, there is still no consensus nor standards for how feature extraction should be performed. E.g. within the field of parcellation based connectivity analysis, it has been shown that the choice of atlas highly influences the final result of the study [89, 143, 179]. A way to overcome the choice of atlas, is to use decomposition methods, such as ICA and MSAA, which find the components that best explain the data. However, there are also challenges for these methods, e.g. they are computationally more expensive, the optimization of the model parameters is a non-convex problem (which means that solutions change with different initializations), and there is so far no gold standard for how to match the components across different datasets.

In our studies, we showed that both ICA and MSAA can extract stable brain networks, even on multi-site data, for which the networks were also stable across datasets when using transfer learning approach 3. Since we in Study 2 found that this transfer learning approach was superior both in its stability, predictive performance, reduced computational complexity and because it enabled a direct coupling across datasets, we consider this a promising solution for using decomposition methods across datasets in future studies.

For the MSAA model, we showed that the extracted networks were comparable to those found by ICA, and that the networks could be used for subsequent predictions, with similar (Study 2–3) and even superior prediction performance (Study 1) . We successfully implemented the spotlight approach, which restricts the seed regions to predefined regions of interest. Furthermore, we showed that computational efficiency of MSAA can be improved by preceding the analysis with PCA, which can be used to substantially reduce the computational and time complexity allowing its application to even larger datasets.

Multi-site data and external validation for predictive modelling

When the fMRI field started using machine learning methods for predictive modelling, it was hoped that the use of cross validation would make the findings more robust. However, in the last decade, it has been shown that even predictive modelling studies can be overfitted and that these studies also struggle with reproducibility across datasets [17, 176]. Increasing the sample size has been one of the main suggestions for overcoming the reproducibility challenges; however,

our studies and other recent publications, indicate that "it is not all about larger datasets" [202]. Whereas larger sample sizes in one way increase the statistical power for the analysis, the data from such studies will typically be more heterogenous, because the data is acquired at different sites or because it is collected with less strict inclusion/exclusion criteria. This can both be a benefit, because models that are trained on a more heterogenous group might be more reproducible to other datasets, but it can also be challenging if the sampling and measurement biases become larger than the targeted biological signal [77]. These factors should thus be carefully considered in relation to the question of interest for the study.

In Studies 2–3 we used intra-site cross validation to develop our models on a multi-site discovery dataset, and tested the reproducibility on an external dataset. Throughout our analyses, we opted towards using simple and robust steps rather than optimizing the performance on the discovery dataset, with the goal to find reproducible biomarkers that would generalize towards the external test dataset.

Using this approach, we were able to classify the diagnosis of the participants with high and similar performance on both the discovery and the external test data. On the contrary, predictions on the PANSS scales were low to moderate on the discovery data and generally did not reproduce to the external test data. One potential explanation for the low performances of the PANSS prediction could be that the amount of additional heterogeneity (multi-site data pooled from different sites) was too large in comparison to the signal of interest (PANSS related variation) in the data. More specifically, even though our sample size was relatively large compared to earlier studies, since our data came from multiple sites, we only had data from 19–55 patients from each site. This is one example of where the additional heterogeneity might outweigh the benefits when using multi-site data. However, as discussed in Paper C, there are also other possible explanations for this finding.

Predictions on individual RSNs

We also want to highlight another interesting observation which we consistently found (both for the classification and particularly for the PANSS predictions) when performing individual predictions of the RSNs from the decomposition methods.

First of all, the highest classification was always found for the ensemble decision model. A similar result was found for the parcellation based connectivity analysis, where we investigated the contributions of individual RSNs both through weightmap contributions and individual subpart predictions. This finding shows that it, for all predictions, was an advantage to use information from all RSNs, and that there was no 'single best' RSN. This supports earlier findings that schizophrenia affects a wide range of brain networks.

Secondly, we found that the RSNs that yielded the highest prediction performance on the discovery data did not obtain the best results on external test data. This finding relates to the "multiple comparison paradox" described by Marek et al. who found that correcting for multiple comparisons (and thereby choosing the solution with the highest performance) reduced the probability of successfully replicating results of brain-wise association studies[71]. This result highlights the importance of validating potential biomarkers on external data, instead of 'just'

reporting the solution that obtained the highest prediction performance, presuming that this solution would also obtain the best performance on other datasets, and hopefully even to a broader population.

fMRI biomarkers for subtyping

In the last decade, there has been a large focus on finding more mechanistic disease definitions for psychiatric disorders, as highlighted by initiatives such as RDoC [4]. In this spirit, clustering of fMRI data has the potential to discover disease subtypes with a more homogeneous biology.

In Study 4, we used the multiple co-clustering algorithm to search for disease-related subtypes using connectivity data from fMRI. A major part of our analysis was focused on determining the stability of the clustering, both in relation to different initializations and changes in the dataset (atlases, datasplits and external data). Finally, we used the the most stable clustering solution to search for subtypes with a significant disease association.

Overall, we found that the clustering method was very dependent on variability in the dataset, as well as on the size of the input data (different atlases). Even for the Allen atlas, which only includes 28 RSNs, we found subject clustering was only moderate across initializations. Neverthe-less, using this atlas, we found a subtyping solution, which had a significant diagnosis association both on the discovery and external test dataset. In this solution, there were three potential schizophrenia-related subtypes, which included a predominance of patients with schizophrenia, including all patients with a moderate to high total PANSS score. Furthermore, we found that feature cluster 3 of this subject clustering solution, showed a reproducible linear trend in the connectivity values, such that subject clusters with a predominance of healthy controls had positive connectivity values, whereas subject clusters with a predominance of schziophrenia patients showed negative connectivity. All six connectivity features that were included in this feature cluster were related to the basal ganglia RSN, indicating the importance of subcortical regions. Since the Allen atlas only included one basal ganglia RSN, and no further subdivision into individual subcortical regions, we were not able to determine if this finding was specifically related to activity changes in the striatum.

Based on our findings, we believe that fMRI based clustering methods, such as MCC, have a great potential towards finding new disease related subtypes with a more homogeneous biology. However, as for the predictive modeling studies, we still think there are important methodological questions that need further investigations, particularly in relation to the stability of the methods.

In Study 4, we have performed the subtyping analysis directly on the fMRI data, which provides the opportunity find completely data-driven disease subtypes. Alternatively, the clustering could also be performed on combined datatypes (polytopic learning), such as combining the analysis of fMRI connectivity features and data from clinical scales or other (neuroimaging) biomarker modalities [14]. Finally, fMRI data can also be used to support subtypes that are defined based on other biomarker outputs or clinical scales. E.g. Chen et al. recently used a large multi-site PANSS dataset to search for a new factorization of the PANSS scale, and found a 4-factor model with improved consistency compared to the traditional three subscales [50, 203]. This PANSS factorization (also referred to as psychopathological subtype) was found solely on data from the

PANSS scales, but the authors subsequently used fMRI to study the neurobiological foundation of their findings [203].

6.1   **Future directions**

Looking back on our four studies, we believe that one of the most promising potentials of fMRI biomarkers are within the field of disease subtyping, either by using clustering directly on the fMRI data (as in Study 4) or by using fMRI to support subtypes found on other datatypes. There are still important methodological challenges that need to be addressed, but the discovery of stable and reproducible subtypes within or even between psychiatric disorders, would have great clinical potential for many applications, such as better diagnostics, more efficient treatment opportunities and personalized medicine [14, 129].

If we were to continue our research for another year, we would thus focus our efforts on further explorations of the subtyping methods. Firstly we would like to continue our work on mapping how the stability is affected by different factors. For example, it would be interesting to determine how the stability depends on the iterations (and thereby convergence) and the number of features in the input space, rather than 'just' comparing two atlases as done in Study 4. Furthermore, we would be interested in further exploration of the potential of polytopic learning, and combine the connectivity data with other available information, such as the individual items of the PANSS scores where these are available. Finally, the authors of the MCC algorithm have also developed another clustering method specifically for fMRI connectivity data, which keeps the structure of the connectivity matrix without vectorization[128, 204]. This method, however, comes with other limitations (as described in Paper D), and we think it would thus be interesting to further investigate their models and to develop ways to overcome the limitations imposed by the additional constraints.

With regards to future predictive modelling studies, we also believe that there are substantial reasons to be optimistic, and that the combination of machine learning and multi-site data carries great potential [15]. We have shown that it is already possible to perform multi-site classifications on diagnostic labels, which have high performance and reproduce across datasets, demonstrating the potential of using multi-site data and rigorous prediction frameworks, including data-driven feature extraction methods. We hope that data sharing initiatives will continue to expand to include even more data from patients in various diseases, and hopefully also more phenotypic measures, by which it would be possible to study more fine-grained trends in the the future.

We also believe that it will be necessary to perform more structured investigations to explore the test-retest variability of fMRI data. This includes developing standards and aligning best practices for acquisition setups (e.g. scanner settings and calibration, but also proper training of technical personal), experimental paradigms and data analysis. The required reliability might vary depending on the question of interest. E.g. if a fMRI paradigm has a low within-subject reproducibility, it is unlikely to have much utility for supporting clinical decisions, no matter how advanced machine learning method are used to aim for individual predictions. However, if the paradigm has a sufficient group-level reproducibility (which can be good, even for tasks with a low

within-subject reproducibility[205]), it might still be suitable to use this task for parallel-group designs in drug development, and might thus be considered in cases where the task has sufficient sensitivity [18, 206].

Studies like ours are important towards solving methodological challenges, but we want to highlight that different activities are also needed for the clinical translation of fMRI biomarkers. For example, even if a fMRI biomarker is found to reliably obtain high prediction performances across datasets, often real-world validation is still needed, because the datasets which are collected for research purposes do not necessarily reflect the true clinical population[1][207]. Furthermore, for a biomarker to be included into clinical practice, it should be feasible to implement it in a real-world setting (e.g. not be too costly or time-consuming), be safe and the methods needed should be 'accepted' by the community [94]. This is particularly important in the field of machine learning, where activities are still ongoing to determine how it can, and should, be used to support healthcare [208, 209] and drug development [210, 211].

6.2 Conclusion

In conclusion, we see the contributions of our work as important methodological steps towards using machine learning and multi-site data to find robust and reproducible fMRI biomarkers. Even though important methodological challenges are yet to be resolved, we are optimistic about the future of the field moving towards applications with increasing clinical utility. We hope that data sharing initiatives will keep on growing, both in size and with respect to the kind of data that are included, such that it will be possible to draw firm conclusions on specific phenotypic measures in the future. Finally, although important contributions are being made on many fronts, we believe that the field of subtyping in particular has a considerable potential to change the clinical utility of future fMRI biomarkers.

---

[1]Patients that take part in research studies, particularly if these have a rather high patient burden, might have higher functioning and lower disease severity and less comorbidity

BIBLIOGRAPHY

[1] V L Feigin, T Vos, E Nichols, M O Owolabi, W M Carroll, M Dichgans, G Deuschl, P Parmar, M Brainin, and C Murray. The global burden of neurological disorders: translating evidence into policy. *Lancet Neurol*, 19:255–265, 2020. 1474-4465 Feigin, Valery L Vos, Theo Nichols, Emma Owolabi, Mayowa O Carroll, William M Dichgans, Martin Deuschl, Günther Parmar, Priya Brainin, Michael Murray, Christopher Journal Article Research Support, Non-U.S. Gov't Review England Lancet Neurol. 2020 Mar;19(3):255-265. doi: 10.1016/S1474-4422(19)30411-9. Epub 2019 Dec 5.

[2] Henry Markram. Seven challenges for neuroscience. *Functional neurology*, 28:145–151, 2013.

[3] Association American Psychiatric. *Diagnostic and Statistical Manual of Mental Disorders, DSM-5*. American Psychiatric Association, 2013.

[4] T Insel, B Cuthbert, M Garvey, R Heinssen, D S Pine, K Quinn, C Sanislow, and P Wang. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders. *Am J Psychiatry*, 167:748–751, 2010. 1535-7228 Insel, Thomas Cuthbert, Bruce Garvey, Marjorie Heinssen, Robert Pine, Daniel S Quinn, Kevin Sanislow, Charles Wang, Philip Journal Article United States Am J Psychiatry. 2010 Jul;167(7):748-51. doi: 10.1176/appi.ajp.2010.09091379.

[5] Ana Vilar, Víctor Pérez-Sola, María Jesús Blasco, Elena Pérez-Gallo, Laura Ballester Coma, Santiago Batlle Vila, Jordi Alonso, Antoni Serrano-Blanco, and Carlos G Forero. Translational research in psychiatry: The research domain criteria project (rdoc). *Revista de Psiquiatría y Salud Mental (English Edition)*, 12:187–195, 2019.

[6] N V Kraguljac, W M McDonald, A S Widge, C I Rodriguez, M Tohen, and C B Nemeroff. Neuroimaging biomarkers in schizophrenia. *Am J Psychiatry*, 178:509–521, 2021.

[7] FDA-NIH. Best (biomarkers, endpoints, and other tools) resource, 2016.

[8] S. R. Kay, A. Fiszbein, and L. A. Opler. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophr Bull*, 13(2):261–76, 1987. Kay, S R Fiszbein, A Opler, L A Journal Article United States Schizophr Bull. 1987;13(2):261-76. doi: 10.1093/schbul/13.2.261.

[9] Christoph U Correll and Nina R Schooler. Negative symptoms in schizophrenia: a review and clinical guide for recognition, assessment, and treatment. *Neuropsychiatric disease and treatment*, pages 519–534, 2020.

[10] Silvana Galderisi, Armida Mucci, Robert W Buchanan, and Celso Arango. Negative symptoms of schizophrenia: new developments and unanswered research questions. *The Lancet Psychiatry*, 5(8):664–677, 2018.

[11] Koichi Kaneko. Negative symptoms and cognitive impairments in schizophrenia: two key symptoms negatively influencing social functioning. *Yonago acta medica*, 61(2):091–102, 2018.

[12] Evangelos Papanastasiou and Sukhwinder S Shergill. Why should pharmacological trials in schizophrenia employ functional magnetic resonance imaging (fmri)? *Journal of Psychopharmacology*, 35(9):1158–1160, 2021. PMID: 33908311.

[13] Debo Dong, Yulin Wang, Xuebin Chang, Cheng Luo, and Dezhong Yao. Dysfunction of large-scale brain networks in schizophrenia: A meta-analysis of resting-state functional connectivity. *Schizophrenia Bulletin*, 44:168–181, 2018.

[14] L Miranda, R Paul, B Pütz, N Koutsouleris, and B Müller-Myhsok. Systematic review of functional mri applications for psychiatric disease subtyping. *Front Psychiatry*, 12:665536, 2021. 1664-0640 Miranda, Lucas Paul, Riya Pütz, Benno Koutsouleris, Nikolaos Müller-Myhsok, Bertram Systematic Review Switzerland Front Psychiatry. 2021 Oct 22;12:665536. doi: 10.3389/fpsyt.2021.665536. eCollection 2021.

[15] V D Calhoun, G D Pearlson, and J Sui. Data-driven approaches to neuroimaging biomarkers for neurological and psychiatric disorders: emerging approaches and examples. *Curr Opin Neurol*, 34:469–479, 2021.

[16] Jing Sui, Rongtao Jiang, Juan Bustillo, and Vince Calhoun. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biological psychiatry*, 88(11):818–828, 2020.

[17] Choong-Wan Woo, Luke J Chang, Martin A Lindquist, and Tor D Wager. Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, 20:365–377, 2017.

[18] Owen Carmichael, Adam J. Schwarz, Christopher H. Chatham, David Scott, Jessica A. Turner, Jaymin Upadhyay, Alexandre Coimbra, James A. Goodman, Richard Baumgartner, Brett A. English, John W. Apolzan, Preetham Shankapal, and Keely R. Hawkins. The role of fmri in drug development. *Drug Discovery Today*, 23(2):333–348, 2018.

[19] Saori C Tanaka, Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, Naohiro Okada, Ryuichiro Hashimoto, Go Okada, Yuki Sakai, Jun Morimoto, Jin Narumoto, Yasuhiro Shimada, Hiroaki Mano, Wako Yoshida, Ben Seymour, Takeshi Shimizu, Koichi Hosomi, Youichi Saitoh, Kiyoto Kasai, Nobumasa Kato, Hidehiko Takahashi, Yasumasa Okamoto, Okito Yamashita, Mitsuo Kawato, and Hiroshi Imamizu. A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data*, 8, 2021.

[20] The Center for Biomedical Research Excellence (COBRE).

[21] H G Schnack. Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophr Res*, 214:34–42, 2019. 1573-2509 Schnack, Hugo G Journal Article Review Netherlands Schizophr Res. 2019 Dec;214:34-42. doi: 10.1016/j.schres.2017.10.023. Epub 2017 Nov 1.

[22] Institute of health Metrics and Evaluation (IHME). Global Health Data Exchange (GHDx).Accessed 25 September 2021.

[23] fact sheet. Assesed 10th of january 2023 World health organization (WHO).

[24] René S. Kahn, Iris E. Sommer, Robin M. Murray, Andreas Meyer-Lindenberg, Daniel R. Weinberger, Tyrone D. Cannon, Michael O'Donovan, Christoph U. Correll, John M. Kane, Jim Van Os, and et al. Schizophrenia. *Nature Reviews Disease Primers*, 1(1):15067, 2015.

[25] David J. Castle and Robin M. Murray. The neurodevelopmental basis of sex differences in schizophrenia. *Psychological Medicine*, 21(3):565–575, 1991.

[26] Robin M Murray, Vishal Bhavsar, Giada Tripoli, and Oliver Howes. 30 years on: How the neurodevelopmental hypothesis of schizophrenia morphed into the developmental risk factor model of psychosis. *Schizophrenia Bulletin*, 43(6):1190–1196, 2017.

[27] Alastair G. Cardno and Irving I. Gottesman. Twin studies of schizophrenia: From bow-and-arrow concordances to star wars mx and functional genomics. *American Journal of Medical Genetics*, 97(1):12–17, 2000.

[28] Mads G. Henriksen, Julie Nordgaard, and Lennart B. Jansson. Genetics of schizophrenia: Overview of methods, findings and limitations. *Frontiers in Human Neuroscience*, 11, 2017.

[29] Robert A McCutcheon, Tiago Reis Marques, and Oliver D Howes. Schizophrenia—an overview. *JAMA psychiatry*, 77(2):201–210, 2020.

[30] Organization World Health. Icd-10 : international statistical classification of diseases and related health problems : tenth revision, 2004 2004.

[31] P. D. Harvey. Assessing disability in schizophrenia: tools and contributors. *J Clin Psychiatry*, 75(10):e27, 2014.

[32] A. Reichenberg, C. Feo, D. Prestia, C. R. Bowie, T. L. Patterson, and P. D. Harvey. The course and correlates of everyday functioning in schizophrenia. *Schizophr Res Cogn*, 1(1):e47–e52, 2014.

[33] Oliver D. Howes, Rob Mccutcheon, Ofer Agid, Andrea De Bartolomeis, Nico J.M. Van Beveren, Michael L. Birnbaum, Michael A.P. Bloomfield, Rodrigo A. Bressan, Robert W. Buchanan, William T. Carpenter, and et al. Treatment-resistant schizophrenia: Treatment response and resistance in psychosis (trrip) working group consensus guidelines on diagnosis and terminology. *American Journal of Psychiatry*, 174(3):216–229, 2017.

[34] John M. Kane, Ofer Agid, Marjorie L. Baldwin, Oliver Howes, Jean-Pierre Lindenmayer, Stephen Marder, Mark Olfson, Steven G. Potkin, and Christoph U. Correll. Clinical guidance on the identification and management of treatment-resistant schizophrenia. *The Journal of Clinical Psychiatry*, 80(2), 2019.

[35] Oliver D Howes and Shitij Kapur. The dopamine hypothesis of schizophrenia: version iii—the final common pathway. *Schizophrenia bulletin*, 35(3):549–562, 2009.

[36] Oliver D. Howes and Ekaterina Shatalina. Integrating the neurodevelopmental and dopamine hypotheses of schizophrenia and the role of cortical excitation-inhibition balance. *Biological Psychiatry*, 92(6):501–513, 2022.

[37] Thomas R Insel. Rethinking schizophrenia. *Nature*, 468(7321):187–193, 2010.

[38] Jodi J Weinstein, Muhammad O Chohan, Mark Slifstein, Lawrence S Kegeles, Holly Moore, and Anissa Abi-Dargham. Pathway-specific dopamine abnormalities in schizophrenia. *Biological psychiatry*, 81(1):31–42, 2017.

[39] D Jacobs and T Silverstone. Dextroamphetamine-induced arousal in human subjects as a model for mania. *Psychological medicine*, 16(2):323–329, 1986.

[40] William Pettersson-Yeo, Paul Allen, Stefania Benetti, Philip McGuire, and Andrea Mechelli. Dysconnectivity in schizophrenia: Where are we now? *Neuroscience Biobehavioral Reviews*, 35(5):1110–1124, 2011.

[41] Karl Friston, Harriet R Brown, Jakob Siemerkus, and Klaas E Stephan. The dysconnection hypothesis (2016). *Schizophrenia Research*, 176:83–94, 2016. 27450778[pmid] PMC5147460[pmcid] S0920-9964(16)30331-0[PII].

[42] Martijn P Van Den Heuvel and Alex Fornito. Brain networks in schizophrenia. *Neuropsychology review*, 24:32–48, 2014.

[43] Ulrich Ettinger, Christine Mohr, Diane C Gooding, Alex S Cohen, Alexander Rapp, Corinna Haenschel, and Sohee Park. Cognition and brain function in schizotypy: A selective review. *Schizophrenia Bulletin*, 41:S417–S426, 2015.

[44] Jack J Blanchard, Lindsay M Collins, Minu Aghevli, Winnie W Leung, and Alex S Cohen. Social anhedonia and schizotypy in a community sample: The maryland longitudinal study of schizotypy. *Schizophrenia Bulletin*, 37:587–602, 2011.

[45] Thomas R Kwapil. Social anhedonia as a predictor of the development of schizophrenia-spectrum disorders. *Journal of Abnormal Psychology*, 107:558–565, 1998.

[46] Oliver J Mason. The assessment of schizotypy and its clinical relevance. *Schizophrenia Bulletin*, 41:S374–S385, 2015.

[47] S. R. Kay, A. Fiszbein, and L. A. Opler. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophr Bull*, 13(2):261–76, 1987.

[48] Stanley R. Kay. *Positive and negative syndromes in schizophrenia : assessment and research / Stanley R. Kay.* Brunner/Mazel New York, 1991.

[49] Dwight Dickinson, Danielle N Pratt, Evan J Giangrande, Meilin Grunnagle, Jennifer Orel, Daniel R Weinberger, Joseph H Callicott, and Karen F Berman. Attacking heterogeneity in schizophrenia by deriving clinical subgroups from widely available symptom data. *Schizophrenia Bulletin*, 44:101–113, 2018.

[50] Ji Chen, Kaustubh R Patil, Susanne Weis, Kang Sim, Thomas Nickl-Jockschat, Juan Zhou, André Aleman, Iris E Sommer, Edith J Liemburg, Felix Hoffstaedter, Ute Habel, Birgit Derntl, Xiaojin Liu, Jona M Fischer, Lydia Kogler, Christina Regenbogen, Vaibhav A Diwadkar, Jeffrey A Stanley, Valentin Riedl, Renaud Jardri, Oliver Gruber, Aristeidis Sotiras, Christos Davatzikos, Simon B Eickhoff, Agna A Bartels-Velthuis, Richard Bruggeman, Stynke Castelein, Frederike Jörg, Gerdina H M Pijnenborg, Henderikus Knegtering, and Ellen Visser. Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: An international machine learning study. *Biological Psychiatry*, 87:282–293, 2020.

[51] Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

[52] Kenneth K Kwong, John W Belliveau, David A Chesler, Inna E Goldberg, Robert M Weisskoff, Brigitte P Poncelet, David N Kennedy, Bernice E Hoppel, Mark S Cohen, and Robert Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679, 1992.

[53] Sasitorn Petcharunpaisan, Joana Ramalho, and Mauricio Castillo. Arterial spin labeling in neuroimaging. *World journal of radiology*, 2(10):384–398, 2010.

[54] William Beecher Scoville and Brenda Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, 20(1):11, 1957.

[55] Vince Calhoun, Tulay Adali, Godfrey Pearlson, and James Pekar. A infomax method for performing {ICA} of {fMRI} data in the complex domain, 2002.

[56] C F Beckmann and S M Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23:137–152, 2004.

[57] Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, and Christian F Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106:13040–13045, 2009.

[58] Vince D Calhoun, Jingyu Liu, and Tülay Adali. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *NeuroImage*, 45:163–172, 2009.

[59] Neil D Woodward and Carissa J Cascio. Resting-state functional connectivity in psychiatric disorders. *JAMA psychiatry*, 72(8):743–744, 2015.

[60] Bindu Menon. Towards a new model of understanding – the triple network, psychopathology and the structure of the mind. *Medical Hypotheses*, 133:109385, 2019.

[61] Denise C Park and Ian M McDonough. The dynamic aging mind: Revelations from functional neuroimaging research. *Perspectives on Psychological Science*, 8(1):62–67, 2013.

[62] Patricia A Reuter-Lorenz. Aging and cognitive neuroimaging: A fertile union. *Perspectives on Psychological Science*, 8(1):68–71, 2013.

[63] Nico UF Dosenbach, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361, 2010.

[64] Karsten Specht. Current challenges in translational and clinical fmri and future directions. *Frontiers in psychiatry*, 10:924, 2020.

[65] Scott E Maxwell, Michael Y Lau, and George S Howard. Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487, 2015.

[66] M R Arbabshirani, S Plis, J Sui, and V D Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145:137–165, 2017.

[67] H. G. Schnack and R. S. Kahn. Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Front Psychiatry*, 7:50, 2016.

[68] Darrel A. Regier, William E. Narrow, Diana E. Clarke, Helena C. Kraemer, S. Janet Kuramoto, Emily A. Kuhl, and David J. Kupfer. Dsm-5 field trials in the united states and canada, part ii: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, 170(1):59–70, 2013.

[69] Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics Data Analysis*, 71:52–78, 2014.

[70] Andre F Marquand, Thomas Wolfers, Maarten Mennes, Jan Buitelaar, and Christian F Beckmann. Beyond lumping and splitting: A review of computational approaches for stratifying psychiatric disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1:433–447, 2016.

[71] Scott Marek, Brenden Tervo-Clemmens, Finnegan J Calabro, David F Montez, Benjamin P Kay, Alexander S Hatoum, Meghan Rose Donohue, William Foran, Ryland L Miller, Timothy J Hendrickson, Stephen M Malone, Sridhar Kandala, Eric Feczko, Oscar Miranda-Dominguez, Alice M Graham, Eric A Earl, Anders J Perrone, Michaela Cordova, Olivia Doyle, Lucille A Moore, Gregory M Conan, Johnny Uriarte, Kathy Snider, Benjamin J Lynch, James C Wilgenbusch, Thomas Pengo, Angela Tam, Jianzhong Chen, Dillan J Newbold, Annie Zheng, Nicole A Seider, Andrew N Van, Athanasia Metoki, Roselyne J Chauvin, Timothy O Laumann, Deanna J Greene, Steven E Petersen, Hugh Garavan, Wesley K Thompson, Thomas E Nichols, B T Thomas Yeo, Deanna M Barch, Beatriz Luna, Damien A Fair, and Nico U F Dosenbach. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603:654–660, 2022.

[72] Benjamin O Turner, Erick J Paul, Michael B Miller, and Aron K Barbey. Small sample sizes reduce the replicability of task-based fmri studies. *Communications Biology*, 1(1):62, 2018.

[73] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9, 2017.

[74] Dustin Scheinost, Stephanie Noble, Corey Horien, Abigail S Greene, Evelyn M R Lake, Mehraveh Salehi, Siyuan Gao, Xilin Shen, David O'Connor, Daniel S Barron, Sarah W Yip, Monica D Rosenberg, and R Todd Constable. Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage*, 193:35–45, 2019.

[75] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[76] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[77] A Yamashita, N Yahata, T Itahashi, G Lisi, T Yamada, N Ichikawa, M Takamura, Y Yoshihara, A Kunimatsu, N Okada, H Yamagata, K Matsuo, R Hashimoto, G Okada, Y Sakai, J Morimoto, J Narumoto, Y Shimada, K Kasai, N Kato, H Takahashi, Y Okamoto, S C Tanaka, M Kawato, O Yamashita, and H Imamizu. Harmonization of resting-state functional mri data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS Biol*, 17:e3000042, 2019.

[78] Ayumu Yamashita, Yuki Sakai, Takashi Yamada, Noriaki Yahata, Akira Kunimatsu, Naohiro Okada, Takashi Itahashi, Ryuichiro Hashimoto, Hiroto Mizuta, Naho Ichikawa, Masahiro

Takamura, Go Okada, Hirotaka Yamagata, Kenichiro Harada, Koji Matsuo, Saori C Tanaka, Mitsuo Kawato, Kiyoto Kasai, Nobumasa Kato, Hidehiko Takahashi, Yasumasa Okamoto, Okito Yamashita, and Hiroshi Imamizu. Generalizable brain network markers of major depressive disorder across multiple imaging sites. *PLoS biology*, 18:e3000966, 2020.

[79] Koene RA Van Dijk, Trey Hedden, Archana Venkataraman, Karleyton C Evans, Sara W Lazar, and Randy L Buckner. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *Journal of neurophysiology*, 103(1):297–321, 2010.

[80] Stephanie Noble, Dustin Scheinost, and R Todd Constable. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage*, 203:116157, 2019.

[81] Chloe Hutton, Andreas Bork, Oliver Josephs, Ralf Deichmann, John Ashburner, and Robert Turner. Image distortion correction in fmri: a quantitative evaluation. *Neuroimage*, 16(1):217–240, 2002.

[82] Joshua Carp. On the plurality of (methodological) worlds: estimating the analytic flexibility of fmri experiments. *Frontiers in neuroscience*, 6:149, 2012.

[83] Stephen C Strother. Evaluating fmri preprocessing pipelines. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):27–41, 2006.

[84] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth Dupre, Madeleine Snyder, Hiroyuki Oya, Satrajit S Ghosh, Jessey Wright, Joke Durnez, Russell A Poldrack, and Krzysztof J Gorgolewski. fmriprep: a robust preprocessing pipeline for functional mri. *Nature Methods*, 16:111–116, 2019.

[85] Meichen Yu, Kristin A Linn, Philip A Cook, Mary L Phillips, Melvin McInnis, Maurizio Fava, Madhukar H Trivedi, Myrna M Weissman, Russell T Shinohara, and Yvette I Sheline. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. *Human Brain Mapping*, 39:4213–4227, 2018.

[86] Norihide Maikusa, Yinghan Zhu, Akiko Uematsu, Ayumu Yamashita, Kousaku Saotome, Naohiro Okada, Kiyoto Kasai, Kazuo Okanoya, Okito Yamashita, Saori C Tanaka, et al. Comparison of traveling-subject and combat harmonization methods for assessing structural brain characteristics. *Human brain mapping*, 42(16):5278–5287, 2021.

[87] Rogier A Feis, Stephen M Smith, Nicola Filippini, Gwenaëlle Douaud, Elise GP Dopper, Verena Heise, Aaron J Trachtenberg, John C van Swieten, Mark A van Buchem, Serge ARB Rombouts, et al. Ica-based artifact removal diminishes scan site differences in multi-center resting-state fmri. *Frontiers in neuroscience*, 9:395, 2015.

[88] Huanjie Li, Stephen M Smith, Staci Gruber, Scott E Lukas, Marisa M Silveri, Kevin P Hill, William DS Killgore, and Lisa D Nickerson. Denoising scanner effects from multimodal mri data using linked independent component analysis. *Neuroimage*, 208:116388, 2020.

[89] Alexandre Abraham, Michael P Milham, Adriana Di Martino, R Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, 147:736–745, 2017.

[90] Richard E Latchaw, Mark J Alberts, Michael H Lev, John J Connors, Robert E Harbaugh, Randall T Higashida, Robert Hobson, Chelsea S Kidwell, Walter J Koroshetz, Vincent Mathews, et al. Recommendations for imaging of acute ischemic stroke: a scientific statement from the american heart association. *Stroke*, 40(11):3646–3678, 2009.

[91] Massimo Filippi, Maria A Rocca, Olga Ciccarelli, Nicola De Stefano, Nikos Evangelou, Ludwig Kappos, Alex Rovira, Jaume Sastre-Garriga, Mar Tintorè, Jette L Frederiksen, et al. Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines. *The Lancet Neurology*, 15(3):292–303, 2016.

[92] K Specht, M Scheffler, J Reinartz, and J Reul. Experiences and applicability of presurgical real-time fmri. *Rivista di Neuroradiologia*, 16(6):1092–1096, 2003.

[93] Stefan Knecht, Bianca Dräger, Michael Deppe, Lars Bobe, Hubertus Lohmann, Agnes Flöel, E-B Ringelstein, and Henning Henningsen. Handedness and hemispheric language dominance in healthy humans. *Brain*, 123(12):2512–2518, 2000.

[94] Andrea Mechelli and Sandra Vieira. From models to tools: clinical translation of machine learning studies in psychosis. *npj Schizophrenia*, 6, 2020.

[95] Cristina Scarpazza, Lea Baecker, Sandra Vieira, and Andrea Mechelli. Applications of machine learning to brain disorders. In *Machine learning*, pages 45–65. Elsevier, 2020.

[96] Sandra Vieira, Walter Hugo Lopez Pinaya, and Andrea Mechelli. Introduction to machine learning. In *Machine learning*, pages 1–20. Elsevier, 2020.

[97] Adam J. Schwarz, Lino Becerra, Jaymin Upadhyay, Julie Anderson, Richard Baumgartner, Alexandre Coimbra, Jeff Evelhoch, Richard Hargreaves, Brigitte Robertson, Smriti Iyengar, Johannes Tauscher, David Bleakman, and David Borsook. A procedural framework for good imaging practice in pharmacological fmri studies applied to drug development 1: processes and requirements. *Drug Discovery Today*, 16(13):583–593, 2011.

[98] Adam J. Schwarz, Lino Becerra, Jaymin Upadhyay, Julie Anderson, Richard Baumgartner, Alexander Coimbra, Jeff Evelhoch, Richard Hargreaves, Brigitte Robertson, Smriti Iyengar, Johannes Tauscher, David Bleakman, and David Borsook. A procedural framework for good imaging practice in pharmacological fmri studies applied to drug development 2: protocol optimization and best practices. *Drug Discovery Today*, 16(15):671–682, 2011.

[99] Dean F Wong, Johannes Tauscher, and Gerhard Gründer. The role of imaging in proof of concept for cns drug discovery and development. *Neuropsychopharmacology*, 34(1):187–203, 2009.

[100] Alaleh Sadraee, Martin Paulus, and Hamed Ekhtiari. fmri as an outcome measure in clinical trials: A systematic review in clinicaltrials.gov. *Brain and Behavior*, 11(5), 2021.

[101] E Fuller Torrey, Robert H Yolken, H Richard Lamb, Wendy W Simmons, Elizabeth Sinclair, John Snook, JD Executive Director, and Treatment Advocacy Center. Why nimh should not stop new drug trials for schizophrenia, 2018.

[102] Florian Lasch, Eftychia-Eirini Psarelli, Ralf Herold, Andrea Mattsson, Lorenzo Guizzaro, Frank Pétavy, and Anja Schiel. The impact of covid-19 on the initiation of clinical trials in europe and the united states. *Clinical Pharmacology & Therapeutics*, 111(5):1093–1102, 2022.

[103] Islam R Younis, Mathangi Gopalakrishnan, Mitchell Mathis, Mehul Mehta, Ramana Uppoor, Hao Zhu, and Tiffany Farchione. Association of end point definition and randomized clinical trial duration in clinical trials of schizophrenia medications. *JAMA psychiatry*, 77(10):1064–1071, 2020.

[104] K E Stephan, K J Friston, and C D Frith. Dysconnection in schizophrenia: From abnormal synaptic plasticity to failures of self-monitoring. *Schizophrenia Bulletin*, 35:509–527, 2009.

[105] Gemma Modinos, Remco Renken, Simone G Shamay-Tsoory, Johan Ormel, and Andr Aleman. Neurobiological correlates of theory of mind in psychosis proneness. *Neuropsychologia*, 48:3715–3724, 2010.

[106] Philipp Kanske, Anne Böckler, Fynn Mathis Trautwein, Franca H Parianen Lesemann, and Tania Singer. Are strong empathizers better mentalizers? evidence for independence and interaction between the routes of social cognition. *Social Cognitive and Affective Neuroscience*, 11:1383–1392, 2016.

[107] Yi Wang, W H H Liu, Z Li, X H H Wei, X Q Q Jiang, F L L Geng, L Q Q Zou, S S Y Lui, E F C Cheung, C Pantelis, and R C K Chan. Altered corticostriatal functional connectivity in individuals with high social anhedonia. *Psychological Medicine*, 46:1–11, 2016.

[108] Yi Wang, Hai song Shi, Wen hua Liu, Dong jie Xie, Fu lei Geng, Chao Yan, Ya Wang, Ya hui Xiao, Suzanne H W So, Chui-De Chiu, Patrick W L Leung, Eric F C Cheung, Diane C Gooding, and Raymond C K Chan. Trajectories of schizotypy and their emotional and social functioning: An 18-month follow-up study. *Schizophrenia Research*, 2017.

[109] Annalaura Lagioia, Dimitri Van De Ville, Martin Debbané, François Lazeyras, and Stephan Eliez. Adolescent resting state networks and their associations with schizotypal trait expression. *Frontiers in Systems Neuroscience*, 4:1–12, 2010.

[110] Eleni Zarogianni, Amos J Storkey, Eve C Johnstone, David G C Owens, and Stephen M Lawrie. Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. *Schizophrenia Research*, 181:6–12, 2017.

[111] Svetlana V Shinkareva, Hernando C Ombao, Bradley P Sutton, Aprajita Mohanty, and Gregory A Miller. Classification of functional brain images with a spatio-temporal dissimilarity map. *NeuroImage*, 33(1):63–71, 2006.

[112] Gemma Modinos, William Pettersson-Yeo, Paul Allen, Philip K McGuire, André Aleman, and Andrea Mechelli. Multivariate pattern classification reveals differential brain activation during emotional processing in individuals with psychosis proneness. *NeuroImage*, 59:3033–3041, 2012.

[113] Gemma Modinos, Andrea Mechelli, William Pettersson-Yeo, Paul Allen, Philip McGuire, and Andre Aleman. Pattern classification of brain activation during emotional processing in subclinical depression: psychosis proneness as potential confounding factor. *PeerJ*, 1:e42–e42, 2013.

[114] Pierre Orban, Christian Dansereau, Laurence Desbois, Violaine Mongeau-Pérusse, Charles Édouard Giguère, Hien Nguyen, Adrianna Mendrek, Emmanuel Stip, and Pierre Bellec. Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity. *Schizophrenia Research*, 192:167–171, 2018.

[115] Kristina C. Skåtun, Tobias Kaufmann, Nhat Trung Doan, Dag Alnæs, Aldo Córdova-Palomera, Erik G. Jönsson, Helena Fatouros-Bergman, Lena Flyckt, Ingrid Melle, Ole A. Andreassen, Ingrid Agartz, and Lars T. Westlye. Consistent functional connectivity alterations in schizophrenia spectrum disorder: A multisite study. *Schizophrenia Bulletin*, 43(4):914–924, 2017.

[116] Ling-Li Zeng, Huaning Wang, Panpan Hu, Bo Yang, Weidan Pu, Hui Shen, Xingui Chen, Zhening Liu, Hong Yin, Qingrong Tan, Kai Wang, and Dewen Hu. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity mri. *EBioMedicine*, 30:74–85, 2018.

[117] Siyi Li, Na Hu, Wenjing Zhang, Bo Tao, Jing Dai, Yao Gong, Youguo Tan, Duanfang Cai, and Su Lui. Dysconnectivity of multiple brain networks in schizophrenia: A meta-analysis of resting-state functional connectivity. *Frontiers in Psychiatry*, 10, 2019.

[118] B T Yeo, F M Krienen, J Sepulcre, M R Sabuncu, D Lashkari, M Hollinshead, J L Roffman, J W Smoller, L Zöllei, J R Polimeni, B Fischl, H Liu, and R L Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol*, 106:1125–1165, 2011.

[119] Stefan P Koch, Claudia Hägele, John Dylan Haynes, Andreas Heinz, Florian Schlagenhauf, and Philipp Sterzer. Diagnostic classification of schizophrenia patients on the basis of regional reward-related fmri signal patterns. *PLOS ONE*, 10, 2015.

[120] Hao-Ting Wang, Jonathan Smallwood, Janaina Mourao-Miranda, Cedric Huchuan Xia, Theodore D Satterthwaite, Danielle S Bassett, and Danilo Bzdok. Finding the needle in a

high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage*, 216:116745, 2020.

[121] Yun-Shuang Fan, Liang Li, Yue Peng, Haoru Li, Jing Guo, Meiling Li, Siqi Yang, Meng Yao, Jingping Zhao, Hesheng Liu, Wei Liao, Xiaonan Guo, Shaoqiang Han, Qian Cui, Xujun Duan, Yong Xu, Yan Zhang, and Huafu Chen. Individual-specific functional connectome biomarkers predict schizophrenia positive symptoms during adolescent brain maturation. *Human Brain Mapping*, 42:1475–1484, 2021.

[122] Y S Fan, H Li, J Guo, Y Pang, L Li, M Hu, M Li, C Wang, W Sheng, H Liu, Q Gao, X Chen, X Zong, and H Chen. Tracking positive and negative symptom improvement in first-episode schizophrenia treated with risperidone using individual-level functional connectivity. *Brain Connect*, 12:454–464, 2022. 2158-0022 Fan, Yun-Shuang Orcid: 0000-0003-1997-5089 Li, Haoru Guo, Jing Pang, Yajing Li, Liang Hu, Maolin Li, Meiling Wang, Chong Sheng, Wei Liu, Hesheng Gao, Qing Orcid: 0000-0001-8504-6128 Chen, Xiaogang Zong, Xiaofen Chen, Huafu Journal Article Research Support, Non-U.S. Gov't United States Brain Connect. 2022 Jun;12(5):454-464. doi: 10.1089/brain.2021.0061. Epub 2021 Sep 13.

[123] Su Lui, Wei Deng, Xiaoqi Huang, Lijun Jiang, Xiaohong Ma, Huafu Chen, Tijiang Zhang, Xiuli Li, Dongming Li, Ling Zou, et al. Association of cerebral deficits with clinical symptoms in antipsychotic-naive first-episode schizophrenia: an optimized voxel-based morphometry and resting state functional connectivity study. *American Journal of Psychiatry*, 166(2):196–205, 2009.

[124] Uzma Nawaz, Ivy Lee, Adam Beermann, Shaun Eack, Matcheri Keshavan, and Roscoe Brady. Individual variation in functional brain network topography is linked to schizophrenia symptomatology. *Schizophrenia Bulletin*, 47:180–188, 2021.

[125] Roscoe O Brady Jr, Irene Gonsalvez, Ivy Lee, Dost Öngür, Larry J Seidman, Jeremy D Schmahmann, Shaun M Eack, Matcheri S Keshavan, Alvaro Pascual-Leone, and Mark A Halko. Cerebellar-prefrontal network connectivity and negative symptoms in schizophrenia. *American Journal of Psychiatry*, 176(7):512–520, 2019.

[126] Kay H Brodersen, Lorenz Deserno, Florian Schlagenhauf, Zhihao Lin, Will D Penny, Joachim M Buhmann, and Klaas E Stephan. Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage: Clinical*, 4:98–111, 2014.

[127] Zhi Yang, Yong Xu, Ting Xu, Colin W Hoy, Daniel A Handwerker, Gang Chen, Georg Northoff, Xi-Nian Zuo, and Peter A Bandettini. Brain network informed subject community detection in early-onset schizophrenia. *Scientific Reports*, 4, 2015.

[128] T. Tokuda, O. Yamashita, Y. Sakai, and J. Yoshimoto. Clustering of multiple psychiatric disorders using functional connectivity in the data-driven brain subnetwork. *Front Psychiatry*, 12:683280, 2021. 1664-0640 Tokuda, Tomoki Yamashita, Okito Sakai, Yuki Yoshi-

moto, Junichiro Journal Article Switzerland Front Psychiatry. 2021 Aug 18;12:683280. doi: 10.3389/fpsyt.2021.683280. eCollection 2021.

[129] Shitij Kapur, Anthony G Phillips, and Thomas R Insel. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular psychiatry*, 17(12):1174–1179, 2012.

[130] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[131] Kristoffer H Madsen, Laerke G Krohne, Xin-Lu Cai, Yi Wang, and Raymond C K Chan. Perspectives on machine learning for classification of schizotypy using fmri data. *Schizophrenia Bulletin*, 44:S480–S490, 2018.

[132] John Ashburner. Preparing fmri data for statistical analysis. In *fMRI techniques and protocols*, pages 151–178. Springer, 2009.

[133] Nathan W Churchill, Anita Oder, Hervé Abdi, Fred Tam, Wayne Lee, Christopher Thomas, Jon E Ween, Simon J Graham, and Stephen C Strother. Optimizing preprocessing and analysis pipelines for single-subject fmri. i. standard temporal motion and physiological noise correction methods. *Human Brain Mapping*, 33:609–627, 2012.

[134] Jonathan D Power, Mark Plitt, Prantik Kundu, Peter A Bandettini, and Alex Martin. Temporal interpolation alters motion in fmri scans: Magnitudes and consequences for artifact detection. *PloS one*, 12(9):e0182939, 2017.

[135] Joshua Carp. The secret lives of experiments: methods reporting in the fmri literature. *Neuroimage*, 63(1):289–300, 2012.

[136] K J Friston, a P Holmes, K J Worsley, J P Poline, C D Frith, and R S J Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

[137] E Bullmore, M Brammer, S C Williams, S Rabe-Hesketh, N Janot, A David, J Mellers, R Howard, and P Sham. Statistical methods of estimation and inference for functional mr image analysis. *Magnetic Resonance in Medicine*, 35:261–277, 1996.

[138] Russell A Poldrack and Krzysztof J Gorgolewski. Making big data open: data sharing in neuroimaging. *Nat Neurosci*, 17:1510–1517, 2014.

[139] Thomas Nichols and Satoru Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12:419–446, 2003.

[140] Matthew Brett, Will Penny, and Stefan Kiebel. An introduction to random field theory, 2003.

[141] Nessa V Bryce, John C Flournoy, João F Guassi Moreira, Maya L Rosen, Kelly A Sambook, Patrick Mair, and Katie A McLaughlin. Brain parcellation selection: An overlooked decision point with meaningful effects on individual differences in resting-state functional connectivity. *NeuroImage*, 243:118487, 2021.

[142] Simon B Eickhoff, Bertrand Thirion, Gaël Varoquaux, and Danilo Bzdok. Connectivity-based parcellation: Critique and implications. *Human Brain Mapping*, 36:4771–4792, 2015.

[143] Simon B Eickhoff, B T Thomas Yeo, and Sarah Genon. Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19:672–686, 2018.

[144] Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[145] R A Fisher. On the probable error of a coefficient of correlation deduced from a small sample, 1921.

[146] Francesco Benedetti, Alessandro Bernasconi, Marta Bosia, and Enrico Smeraldi. Functional and structural brain correlates of theory of mind and empathy deficits in schizophrenia. *Schizophrenia Research*, 114:154–160, 2009.

[147] Birgit A Völlm, Alexander NW Taylor, Paul Richardson, Rhiannon Corcoran, John Stirling, Shane McKie, John FW Deakin, and Rebecca Elliott. Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. *Neuroimage*, 29(1):90–98, 2006.

[148] Yi Wang, Wen-hua Liu, Zhi Li, Xin-hua Wei, Xin-qing Jiang, David L Neumann, David HK Shum, Eric FC Cheung, and Raymond CK Chan. Dimensional schizotypy and social cognition: an fmri imaging study. *Frontiers in Behavioral Neuroscience*, 9:133, 2015.

[149] Ahmad Abu-Akel and Simone Shamay-Tsoory. Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia*, 49(11):2971–2984, 2011.

[150] Simone G Shamay-Tsoory, Hagai Harari, Judith Aharon-Peretz, and Yechiel Levkovitz. The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*, 46(5):668–677, 2010.

[151] Benjamin A Seitzman, Caterina Gratton, Scott Marek, Ryan V Raut, Nico U F Dosenbach, Bradley L Schlaggar, Steven E Petersen, and Deanna J Greene. A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum. *NeuroImage*, 206:116290, 2020.

[152] Jonathan D Power, Bradley L Schlaggar, and Steven E Petersen. Studying brain organization via spontaneous fmri signal. *Neuron*, 84:681–696, 2014.

[153] Elena Allen, Erik Erhardt, Eswar Damaraju, William Gruner, Judith Segall, Rogers Silva, Martin Havlicek, Srinivas Rachakonda, Jill Fries, Ravi Kalyanam, Andrew Michael, Arvind Caprihan, Jessica Turner, Tom Eichele, Steven Adelsheim, Angela Bryan, Juan Bustillo, Vincent Clark, Sarah Feldstein Ewing, Francesca Filbey, Corey Ford, Kent Hutchison, Rex Jung, Kent Kiehl, Piyadasa Kodituwakku, Yuko Komesu, Andrew Mayer, Godfrey Pearlson, John Phillips, Joseph Sadek, Michael Stevens, Ursina Teuscher, Robert Thoma, and Vince Calhoun. A baseline for the multivariate comparison of resting-state networks. *Frontiers in Systems Neuroscience*, 5, 2011.

[154] M Khosla, K Jamison, G H Ngo, A Kuceyeski, and M R Sabuncu. Machine learning in resting-state fmri analysis. *Magn Reson Imaging*, 64:101–121, 2019.

[155] M J McKeown, S Makeig, G G Brown, T p Jung, S S Kindermann, a J Bell, and T J Sejnowski. Analysis of f mri data by blind separation. *Human Brain Mapping*, 6:160–188, 1998.

[156] Andrew M Michael, Mathew Anderson, Robyn L Miller, Tülay Adali, and Vince D Calhoun. Preserving subject variability in group fmri analysis: performance evaluation of gica vs. iva. *Frontiers in Systems Neuroscience*, 8:106, 2014.

[157] Christian F Beckmann, Clare E Mackay, Nicola Filippini, and Stephen M Smith. Group comparison of resting-state fmri data using multi-subject ica and dual regression. *NeuroImage*, 47:S148, 2009.

[158] V D Calhoun, T Adali, G D Pearlson, and J J Pekar. A method for making group inferences from functional mri data using independent component analysis. *Hum Brain Mapp*, 14:140–151, 2001.

[159] Yi-Ou Li, Tülay Adalı, and Vince D Calhoun. Estimating the number of independent components for functional magnetic resonance imaging data. *Human Brain Mapping*, 28:1251–1266, 2007.

[160] M Wax and T Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33:387–392, 1985.

[161] Srinivas Rachakonda, Eric Egolf, Nicolle Correa, and Vince Calhoun. Group ica of fmri toolbox (gift) manual. *Dostupnez [cit 2011-11-5]*, 2007.

[162] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[163] Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63, 2012.

[164] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36:338–347–338–347, 1994.

[165] Jesper Løve Hinrich, Sophia Elizabeth Bardenfleth, Rasmus Erbou Røge, Nathan William Churchill, Kristoffer Hougaard Madsen, and Morten Mørup. Archetypal analysis for modeling multisubject fmri data. *IEEE journal of selected topics in signal processing*, 10:1160–1171, 2016.

[166] Laerke Gebser Krohne, Yi Wang, Jesper L Hinrich, Morten Moerup, Raymond C K Chan, and Kristoffer H Madsen. Classification of social anhedonia using temporal and spatial network features from a social cognition fmri task. *Human Brain Mapping*, 40:4965–4981, 2019.

[167] Julian Eggert and Edgar Körner. Sparse coding and nmf. *IEEE International Conference on Neural Networks - Conference Proceedings*, 4:2529–2533, 2004.

[168] Andrew E Reineberg, Jessica R Andrews-Hanna, Brendan E Depue, Naomi P Friedman, and Marie T Banich. Resting-state networks predict individual differences in common and specific aspects of executive function. *NeuroImage*, 104:69–78, 2015.

[169] Allen A Champagne, Nicole S Coverdale, Andrew Ross, Yining Chen, Christopher I Murray, David Dubowitz, and Douglas J Cook. Multi-modal normalization of resting-state using local physiology reduces changes in functional connectivity patterns observed in mtbi patients. *NeuroImage: Clinical*, 26:102204, 2020.

[170] Xin-Lu Cai, Dong-Jie Xie, Kristoffer H Madsen, Yong-Ming Wang, Sophie Alida Bögemann, Eric F C Cheung, Arne Møller, and Raymond C K Chan. Generalizability of machine learning for classification of schizophrenia based on resting-state functional mri data. *Human Brain Mapping*, 41:172–184, 2020.

[171] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[172] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[173] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170, 2017.

[174] Jean-Philippe Fortin, Nicholas Cullen, Yvette I Sheline, Warren D Taylor, Irem Aselcioglu, Philip A Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120, 2018.

[175] Christian Dansereau, Yassine Benhajali, Celine Risterucci, Emilio Merlo Pich, Pierre Orban, Douglas Arnold, and Pierre Bellec. Statistical power and prediction accuracy in multisite resting-state fmri connectivity. *Neuroimage*, 149:220–232, 2017.

[176] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017.

[177] Leo Breiman and Philip Spector. Submodel selection and evaluation in regression. the x-random case. *International statistical review/revue internationale de Statistique*, pages 291–319, 1992.

[178] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.

[179] Nicolas Traut, Katja Heuer, Guillaume Lemaître, Anita Beggiato, David Germanaud, Monique Elmaleh, Alban Bethegnies, Laurent Bonnasse-Gahot, Weidong Cai, Stanislas Chambon, et al. Insights from an autism imaging biomarker challenge: promises and threats to biomarker discovery. *NeuroImage*, 255:119171, 2022.

[180] Gaël Varoquaux and Veronika Cheplygina. How i failed machine learning in medical imaging–shortcomings and recommendations. *arXiv preprint arXiv:2103.10292*, 2021.

[181] Delaram Sadeghi, Afshin Shoeibi, Navid Ghassemi, Parisa Moridian, Ali Khadem, Roohallah Alizadehsani, Mohammad Teshnehlab, Juan M Gorriz, Fahime Khozeimeh, Yu-Dong Zhang, Saeid Nahavandi, and U Rajendra Acharya. An overview of artificial intelligence techniques for diagnosis of schizophrenia based on magnetic resonance imaging modalities: Methods, challenges, and future works. *Computers in Biology and Medicine*, 146:105554, 2022.

[182] Luca Steardo Jr, Elvira Anna Carbone, Renato De Filippis, Claudia Pisanu, Cristina Segura-Garcia, Alessio Squassina, Pasquale De Fazio, and Luca Steardo. Application of support vector machine on fmri data as biomarkers in schizophrenia diagnosis: a systematic review. *Frontiers in Psychiatry*, 11:588, 2020.

[183] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[184] Vladimir Vapnik, Steven Golowich, and Alex Smola. Support vector method for function approximation, regression estimation and signal processing. 9, 1996.

[185] Bernhard Scholkopf and Alex J Smola. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, 2002.

[186] Christopher M Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., 2006.

[187] Xilin Shen, Emily S Finn, Dustin Scheinost, Monica D Rosenberg, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nature Protocols*, 12:506–518, 2017.

[188] José Melo. Gaussian processes for regression: a tutorial. *Technical Report*, 2012.

[189] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2004.

[190] Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, and Agustín Mayo-Iscar. A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2):89–109, 2010.

[191] Tomoki Tokuda, Junichiro Yoshimoto, Yu Shimizu, Go Okada, Masahiro Takamura, Yasumasa Okamoto, Shigeto Yamawaki, and Kenji Doya. Multiple co-clustering based on nonparametric mixture models with heterogeneous marginal distributions. *PloS one*, 12(10):e0186566, 2017.

[192] Tomoki Tokuda, Junichiro Yoshimoto, Yu Shimizu, Go Okada, Masahiro Takamura, Yasumasa Okamoto, Shigeto Yamawaki, and Kenji Doya. Identification of depression subtypes and relevant brain regions using a data-driven approach. *Scientific Reports*, 8, 2018.

[193] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[194] *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY, 2008.

[195] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[196] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15:1–25, 2002.

[197] Yujiro Yoshihara, Giuseppe Lisi, Noriaki Yahata, Junya Fujino, Yukiko Matsumoto, Jun Miyata, Gen-Ichi Sugihara, Shin-Ichi Urayama, Manabu Kubota, Masahiro Yamashita, Ryuichiro Hashimoto, Naho Ichikawa, Weipke Cahn, Neeltje E M Van Haren, Susumu Mori, Yasumasa Okamoto, Kiyoto Kasai, Nobumasa Kato, Hiroshi Imamizu, René S Kahn, Akira Sawa, Mitsuo Kawato, Toshiya Murai, Jun Morimoto, and Hidehiko Takahashi. Overlapping but asymmetrical relationships between schizophrenia and autism revealed by brain connectivity. *Schizophrenia Bulletin*, 46:1210–1218, 2020.

[198] H Li, R C K Chan, G M McAlonan, and Q Y Gong. Facial emotion processing in schizophrenia: A meta-analysis of functional neuroimaging data. *Schizophrenia Bulletin*, 36:1029–1039, 2010.

[199] Stefan Haufe, Frank Meinecke, Kai Goergen, Sven Dohne, John Dylan Haynes, Benjamin Blankertz, and Felix Biemann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014.

[200] Du Lei, Walter H L Pinaya, Jonathan Young, Therese Amelsvoort, Machteld Marcelis, Gary Donohoe, David O Mothersill, Aiden Corvin, Sandra Vieira, Xiaoqi Huang, Su Lui, Cristina Scarpazza, Celso Arango, Ed Bullmore, Qiyong Gong, Philip McGuire, and Andrea Mechelli. Integrating machining learning and multimodal neuroimaging to detect schizophrenia at the level of the individual. *Human Brain Mapping*, 41:1119–1135, 2020.

[201] Stefan Leucht, John M. Kane, Werner Kissling, Johannes Hamann, Eva Etschel, and Rolf R. Engel. What does the panss mean? *Schizophrenia Research*, 79(2):231–238, 2005.

[202] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5, 2022.

[203] Ji Chen, Tobias Wensing, Felix Hoffstaedter, Edna C Cieslik, Veronika I Müller, Kaustubh R Patil, André Aleman, Birgit Derntl, Oliver Gruber, Renaud Jardri, Lydia Kogler, Iris E Sommer, Simon B Eickhoff, and Thomas Nickl-Jockschat. Neurobiological substrates of the positive formal thought disorder in schizophrenia revealed by seed connectome-based predictive modeling. *NeuroImage: Clinical*, 30:102666, 2021.

[204] Tomoki Tokuda, Okito Yamashita, and Junichiro Yoshimoto. Multiple clustering for identifying subject clusters and brain sub-networks using functional connectivity matrices without vectorization. *Neural Networks*, 142:269–287, 2021.

[205] Craig M Bennett and Michael B Miller. How reliable are the results from functional magnetic resonance imaging? *Annals of the new York Academy of Sciences*, 1191(1):133–155, 2010.

[206] Pauline Delaveau, Maritza Jabourian, Cédric Lemogne, Sophie Guionnet, Loretxu Bergouignan, and Philippe Fossati. Brain effects of antidepressants in major depression: a meta-analysis of emotional processing studies. *Journal of affective disorders*, 130(1-2):66–74, 2011.

[207] Rashmi Patel, Sherifat Oduola, Felicity Callard, Til Wykes, Matthew Broadbent, Robert Stewart, Thomas KJ Craig, and Philip McGuire. What proportion of patients with psychosis is willing to take part in research? a mental health electronic case register analysis. *BMJ open*, 7(3):e013113, 2017.

[208] World Health Organization et al. Ethics and governance of artificial intelligence for health: Who guidance. 2021.

[209] Nithesh Naik, BM Hameed, Dasharathraj K Shetty, Dishant Swain, Milap Shah, Rahul Paul, Kaivalya Aggarwal, Sufyan Ibrahim, Vathsala Patil, Komal Smriti, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Frontiers in surgery*, page 266, 2022.

[210] Food and Drug Administration. Artificial intelligence/machine learning (ai/ml) based software as a medical device (samd) action plan. 2021.

[211] Philip A Hines, Richard H Guy, Anthony J Humphreys, and Marisa Papaluca-Amati. The european medicines agency's goals for regulatory science to 2025. *Nature reviews Drug discovery*, 18(6):403–404, 2019.

[212] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

Appendices

APPENDIX

<div style="text-align: right;">A</div>

This chapter includes additional specifications and results and Figures that were not included in the main thesis.

A.1 Additional specifications to Study 1

None of the data used in this thesis was collected as part of this PhD project. This section includes additional MRI specifications for the datasets used in the four studies.

A.1.1 Brain atlases used in Study 1

Location of the ROIs used in the literature study atlas (top) and Pooled condition atlas (bottom).

Figure A.1: Center coordinates used in Study 1. Panel A) Literature study coordinates, abbreviations: Amyg, amygdala; AngG, angular gyrus; d/v ACC, dorsal/ventral anterior cingulate cortex; d/v mPFC, dorsal/ventral medial prefrontal cortex; d/v Stri, dorsal/ventral striatum; IPL, inferior parietal lobule; i/dL PFC, inferior/dorsolateral prefrontal cortex; Pcun, precuneus; STS, superior temporal sulcus; TP, temporal pole; TPJ, temporoparietal junction Panel B) Pooled condition coordinates; The MNI coordinates are listed in Supplementary table 1 and 3 of publication B. This Figure is an adated version of the coordinate figure in Paper B.

A.2 Acknowledgement to Visualization tools

Throughout the studies we have used several toolboxes and visualization tools which we would like to thank.

For most brain visualizations, we have used the "BrainNet Viewer" toolbox: https://www.nitrc.org/projects/bnv/ developed by Xia et al. [212]

Furthermore, for 3D brain visualizations in Figure 5.1 we used the VitLam toolbox developed by Hinrich et al [165]. The code for this toolbox can be found via Github: https://github.com/JesperLH/VITLAM.

A.3   Datasets used in Studies 2–4

Number of participants and MRI specifications for each site for datasets 2-4 (D2-4). Much of the avalible information was given by the data descriptor by Tanaka et al from 2021 [19].

A.3.1   Acknowledgement to Multi-site datasets

We would like to thank all investigators and participants who from the data that we used in Studies 2–4. Here we used data from the the DecNef Project Brain Data Repository (https://bicr-resource.atr.jp/srpbsopen/), which was collected as part of the Japanese Strategic Research Program for the Promotion of Brain Science (SRPBS) supported by the Japanese Advanced Research and Development Programs for Medical Innovation (AMED).

The second dataset was from the Mind Research Network and the University of New Mexico funded by a National Institute of Health Center of Biomedical Research Excellence (COBRE) grant [20].

	Discovery (D2-4a)							Test (D2-4b)		
Site number	1	2	3	4	5	6	7	8	9	10
Site acronym	HKH	COR*	KTT	UTO*	ATV*	ATT*	CIN	COBRE	SWA*	KUT*
n participants	29	123	121	132	39	13	39	133	120	203
n HC	29	123	75	96	39	13	39	72	101	159
n SZ	0	0	46	36	0	0	0	61	19	44
PANSS scores avaliable	NA	NA	yes	yes	NA	NA	NA	yes	NA	yes
MRI acquistion parameters										
MRI Scanner	SIEMENS Symp	SIEMENS Verio	SIEMENS Trio	GE Discov	SIEMENS Verio	SIEMENS TimTrio	SIEMENS TimTrio	SIEMENS TimTrio	SIEMENS Verio	SIEMENS TimTrio
Magnitic field strength	3T	3T	3T	3T	3T	3T	3T	3T	3T	3T
Number of channels per coil	head-12ch	head-12ch	head- 8ch	head-24ch	head-12ch	head-12ch	head-12ch	head-12ch	head-12ch	head-32ch
TR (s)	2.7	2.5	2	2.5	2.5	2.5	2.5	2	2.5	2.5
TE (ms)	31	30	30	30	30	30	30	29	30	30
Flip angel (deg)	90	80	90	80	80	80	80	75	80	80
Phase encoding	AP	AP	AP	PA	PA	PA	AP	PA	PA	PA
Matrix	64 x 64	64 x 64	64 x 48	64 x 64	64 x 64	64 x 64	64 x 64	64 x 64	64 x 64	64 x 64
Field of view (mm)	192	212	256 x192	212	212	212	212	240	212	212 x 212
In-plane resolution (mm)	3.0 x 3.0	3.3 x 3.3	4.0 x 4.0	3.3	3.3 x 3.3	3.3 x 3.3	3.3 x 3.3	3.75 x 3.75	3.3 x 3.3	3.3125 X 3.3125
Slice thickness (mm)	3	3.2	4	3.2	3.2	3.2	3.2	3.5	3.2	3.2
Slice gap (mm)	0	0.8	0	0.8	0.8	0.8	0.8	1.05	0.8	0.8
Number of slices	38	40	30	40	39	39 or 40	41	33	40	40
Slice acquisition order	Ascending (IL)	NA	Ascending (IL)	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending
Number of volumes	107 + 5 (dummy)	240 + 4 (dummy)	182	240 + 4 (dummy)	240 + 4 (dummy)	240 + 4 (dummy)	240 + 4 (dummy)	149 + 1 (dummy)	240 + 4 (dummy)	240 + 4 (dummy)
Total scan time	~5 min	10 min	6 min	10 min	10 min	10 mins	10 min	~5 min	10min	10 min
Eye closed/fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
Field map:										
E Echo spacing	NA	0.0005	NA	0.00029	0.00049	0.00049	NA	NA	0.0005	0.00056
Echo time 1	NA	0.00492	NA	0.0049	0.00492	0.00492	NA	NA	0.00492	0.00492
Echo time 2	NA	0.00738	NA	0.0074	0.00738	0.00738	NA	NA	0.00738	0.00738
Blipdir	NA	j-	NA	j	j	j	NA	NA	j	j

Table A.1: Number of participants and MRI acquisition parameters for each site (D2-D4) This table includes the site acronym (HKH, Hiroshima Kajikawa Hospital; COI, Center of Innovation at Hiroshima university; KTT, Kyoto University (Trio); UTO, University of Tokyo Hospital; ATT, Brain Activity Imaging Center ATR-Promotions Inc., Kyoto (Trio); ATV, Brain Activity Imaging Center ATR-Promotions Inc., Kyoto (Verio); CIN, Center for Information and Neural Networks; COBRE, The Center for Biomedical Research Excellence ; SWA, Showa university; KUT, Kyoto University (TimTrio)) and number of participants for each site. The bottom indicates the MRI acquisition parameters.The asterisk, * indicates sites for which participants are scanned with a unified protocol [19, 77]

A.4   **Search on ClinicalTrials.gov**

This section includes the search results for our search for trials in Schizophrenia that have used fMRI. Search words were: "Schizophrenia" [Condition or disease], "fmri OR 'functional MRI' OR 'functional magnetic resonance imaging" [other terms] and "Industry" [Funding type]. We have chosen to exclude trials were the phase was listed as "not applicable" as well as trials that were terminated or not performed in patients with Schizophrenia (this was the case for the last trial on the list).

   Some characteristics from this search are summarized in Figure 2.5

ClinicalTrials.gov Search Results 02/06/2023

	Title	Status	Study Results	Conditions	Interventions	Locations
1	Monotherapy Brexpiprazole (OPC-34712) Trial in the Treatment of Adults With Schizophrenia With Impulsivity	Completed	Has Results	•Schizophrenia With Impulsivity	•Drug: Brexpiprazole	•University of California at Irvine Medical Center, Orange, California, United States
2	The Effects of Aripiprazole on the Processing of Rewards in Schizophrenia	Terminated	Has Results	•Schizophrenia	•Other: fMRI •Drug: Aripiprazole	•Atlanta VA Medical Center, Decatur, Georgia, United States
3	The Effect of Ketamine on Attentiveness	Completed	No Results Available	•Schizophrenia	•Drug: Placebo •Drug: ketamine	
4	Roflumilast Plus Antipsychotics Proof of Mechanism Study in Schizophrenia	Completed	Has Results	•Schizophrenia	•Drug: Roflumilast •Drug: Placebo •Drug: Second generation antipsychotic	•Denmark Hill, London, United Kingdom
5	Modulation of Regional Brain Activation in Schizophrenic Patients by Pharmacological Therapy	Terminated	No Results Available	•Schizophrenia	•Drug: Amisulpride •Drug: Olanzapine •Drug: Haloperidol	
6	Study Assessing SEP-363856 in Male and Female Volunteers With High or Low Schizotype Characteristics	Completed	No Results Available	•Schizophrenia	•Drug: SEP-363856 •Drug: Amisulpride •Drug: Placebo	•Department of Psychiatry, University of Oxford, Warneford Hospital, Headington, Oxford, United Kingdom •University of Manchester, Neuroscience and Psychiatry Unit, Manchester, United Kingdom
7	Pharmacodynamic/Pharmacokinetic Study of AQW051 in Schizophrenia	Completed	No Results Available	•Schizophrenia	•Drug: AQW051 •Drug: Placebo	•West LA VA Healthcare Center (UCLA), Los Angeles, California, United States •Department of Psychiatry & Behavioural Sciences, Feinberg School of Medicine (Northwestern University), Chicago, Illinois, United States •Maryland Psychiatric Research Centre, Spring Grove Hospital Grounds, Baltimore, Maryland, United States •Massachusetts General Hospital (Freedom Trail Clinic), Boston, Massachusetts, United States •Washington University, Saint Louis, Missouri, United States •Columbia University, New York, New York, United States •JUH Clinical Research (Duke University),, Butner, North Carolina, United States
8	A D2 Receptor Occupancy and fMRI Study in Schizophrenic Subjects Treated With Lurasidone	Completed	No Results Available	•Schizophrenia	•Drug: Lurasidone 80 mg •Drug: Lurasidone 120 mg •Drug: Lurasidone 160 mg	•UCI Medical Center, Orange, California, United States
9	Contrasting the Brain Effects of Risperidone and Invega With Functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET) Scanning	Unknown status	No Results Available	•Schizophrenia	•Drug: Risperidone •Drug: Paliperidone	•UC Irvine, Irvine, California, United States
10	Pilot Study of Atomoxetine To Enhance COgnition In Patients With Schizophrenia	Completed	No Results Available	•Schizophrenia	•Drug: Atomoxetine	•Pilgrim Psychiatric Center, Brentwood, New York, United States •Mount Sinai Hospital, New York, New York, United States
11	Cognitive Improvement With Aripiprazole (Abilify) in Patients With Schizophrenia (BMS)	Terminated	No Results Available	•Schizophrenia •Schizoaffective Disorder	•Drug: aripiprazole	•Psychopharmacology Research Clinic, Shreveport, Louisiana, United States
12	POC Study of Pipamperone Added to Stable Treatment With RIS or PAL in Chronic Schizophrenia	Completed	No Results Available	•Chronic Schizophrenia •Schizoaffective Disorder	•Drug: Pipamperone •Drug: Placebo	•University Psychiatric Institute Sint-Jozef, Kortenberg, Belgium

	Title	Status	Study Results	Conditions	Interventions	Locations
13	Neuromodulation for Schizophrenia	Not yet recruiting	No Results Available	• Schizophrenia	• Device: Non-invasive brainstem modulation device (stimulation Randomized) • Device: Non-invasive brainstem modulation device (stimulation-Open Label)	• Centre for Addiction and Mental Health, Toronto, Ontario, Canada
14	Add On Treatment for Cognitive Deficits in Schizophrenia	Completed	No Results Available	• Schizophrenia	• Drug: PF 03654746 • Other: Placebo	• University of Pennsylvania, Philadelphia, Pennsylvania, United States
15	A Study to Evaluate The Effects of RO5545965 in Participants With Negative Symptoms of Schizophrenia Treated With Antipsychotics	Completed	No Results Available	• Schizophrenia	• Drug: Placebo • Drug: RO5545965	• CNS Network, Garden Grove, California, United States • Parexel California Clinical Trials Medical Group, Glendale, California, United States • St Louis Clinical Trials, Saint Louis, Missouri, United States
16	The Effects AZD8529 on Cognition and Negative Symptoms in Schizophrenics	Completed	No Results Available	• Schizophrenia	• Drug: AZD8529 • Drug: Placebo to match AZD8529	• Research Site, Philadelphia, Pennsylvania, United States
17	The Efficacy and Safety of a Selective Estrogen Receptor Beta Agonist (LY500307) for Negative Symptoms and Cognitive Impairment Associated With Schizophrenia	Terminated	Has Results	• Schizophrenia	• Drug: LY500307 150mg • Drug: LY500307 75mg • Drug: Placebo • Drug: LY500307 25mg	• Indiana University Center for NeuroImaging, Indianapolis, Indiana, United States • IU Biostatistics, Indianapolis, Indiana, United States • Prevention and Recovery Center for Early Psychosis, Indianapolis, Indiana, United States • Larue D Carter Memorial Hospital, Indianapolis, Indiana, United States
18	A Randomized, Double-Blind, Placebo Controlled, Two-Period Cross-Over, Proof of Activity Study to Evaluate the Effects of TAK-041 on Motivational Anhedonia as Add-On to Antipsychotics in Participants With Stable Schizophrenia	Completed	Has Results	• Stable Schizophrenia	• Drug: TAK-041 • Drug: Placebo • Drug: Second Generation Antipsychotics (SGA)	• Kings College London, London, United Kingdom
19	Association of Amisulpride Response in Schizophrenia With Brain Image	Unknown status	No Results Available	• Schizophrenia • Schizophreniform Disorder	• Drug: amisulpride	
20	Appetite Increase in Schizophrenia Patients Treated With Atypical Antipsychotics	Completed	No Results Available	• Schizophrenia	• Drug: Olanzapine	• Centre de recherche Fernand-Seguin, Montréal, Quebec, Canada
21	The CAMPUS Project: Cholinergic Augmentation of Cognitive Deficits in Schizophrenia	Terminated	No Results Available	• Schizophrenia	• Drug: donepezil (5-10 mg/day) • Drug: Placebo	• Center for Neuropsychiatric Schizophrenia Research, University of Copenhagen, Dept. F, Bispebjerg Hospital, Copenhagen NV, Denmark • Dept. of Psychiatry O, Rigshospitalet, Blegdamsvej 9, Copenhagen, Denmark • Psychiatric Center, Glostrup, Glostrup, Denmark • Danish Research Center for Magnetic Resonance Imaging, Hvidovre Hospital, Hvidovre, Denmark
22	A Translational and Neurocomputational Evaluation of a Dopamine Receptor 1 Partial Agonist for Schizophrenia	Recruiting	No Results Available	• Early Course Schizophrenia Spectrum Disorder	• Drug: CVL-562 (PF-06412562) 1 mg • Drug: CVL-562 (PF-06412562) 4 mg • Drug: CVL-562 (PF-06412562) 15 mg • Drug: CVL-562 (PF-06412562) 25 mg • Other: Placebo	• Yale University, New Haven, Connecticut, United States • Columbia University, New York, New York, United States • Stony Brook University, Stony Brook, New York, United States • University of Pennsylvania, Philadelphia, Pennsylvania, United States

	Title	Status	Study Results	Conditions	Interventions	Locations
23	A Study To Examine Safety, Pharmacokinetics, And Pharmacodynamic Of Pf 06412562 In Subjects With Schizophrenia	Completed	Has Results	•Schizophrenia	•Drug: PF-06412562 3mg BID •Drug: PF-06412562 9mg BID •Drug: PF-06412562 45mg BID •Other: Placebo	•Arcadia MRI & Imaging Center, Arcadia, California, United States •California Clinical Trials Medical Group, Glendale, California, United States •Glendale Adventist Medical Center, Glendale, California, United States •Maryland Psychiatric Research Center (MPRC) of the University of Maryland, Baltimore, Maryland, United States •CBH Health, LLC, Gaithersburg, Maryland, United States •Foers Long Term Care Pharmacy LLC, Rockville, Maryland, United States
24	A Placebo-Controlled Study of Physiologic Effects of L-methylfolate in Schizophrenia Patients	Completed	No Results Available	•Schizophrenia	•Other: Placebo •Other: L-methylfolate	•Massachusetts General Hospital (MGH) Schizophrenia Program at Freedom Trail Clinic, Boston, Massachusetts, United States
25	Effects of TAK-063 on Preventing Ketamine-Induced Brain Activity Changes as Well as Psychotic-Like Symptoms in Healthy Male Adults	Completed	Has Results	•Ketamine-Induced Brain Activity Changes •Psychotic-like Symptoms	•Drug: Ketamine •Drug: TAK-063 •Drug: TAK-063 Placebo	•The Brain Institute, University of Utah, Salt Lake City, Utah, United States
26	Neurobiological and Neurocognitive Disturbances in First-episode Schizophrenia	Completed	No Results Available	•Schizophrenia	•Drug: quetiapine	•Neurobiology Research Unit, Rigshospitalet, Copenhagen, Denmark •Center for Neuropsychiatric Schizophrenia Research, University of Copenhagen, Psychiatric Center Glostrup, Glostrup, Denmark •Danish Research Center for Magnetic Resonance Imaging, Hvidovre Hospital, Hvidovre, Denmark
27	Dopaminergic, Functional, Structural, and Cognitive Disturbances in First-episode Schizophrenia	Completed	No Results Available	•Schizophrenia	•Drug: zuclopenthixol •Drug: risperidone	•Dept. of Nuclear Medicine, University of Copenhagen, Bispebjerg Hospital, Copenhagen NV, Denmark •University of Copenhagen, Dept. F, Bispebjerg Hospital, Copenhagen NV, Denmark •University of Copenhagen, Dept. of Psychiatry E, Bispebjerg Hospital, Copenhagen NV, Denmark •Neurobiology Research Unit, University of Copenhagen, Rigshospitalet, Copenhagen, Denmark •Center for Neuropsychiatric Schizophrenia Research, University of Copenhagen, Psychiatric Center Glostrup, Glostrup, Denmark •Danish Research Center for Magnetic Resonance Imaging, Hvidovre Hospital, Hvidovre, Denmark
28	SRC Inhibition as a Potential Target for Parkinson's Disease Psychosis	Unknown status	No Results Available	•Parkinson Disease Psychosis	•Drug: Saracatinib •Drug: Placebo Oral Tablet	•Mitul Mehta, London, Camberwell, United Kingdom

	Title	Status	Study Results	Conditions	Interventions	Locations
29	Effects on Social and Cognition Functions of Blonanserin in First Episode Schizophrenia Patients	Unknown status	No Results Available	•First Episode Schizophrenia •Social Function •Cognition Function •Blonanserin	•Drug: Blonanserin •Other: MRI and serum BDNF	•Beijing Huilongguan Hospital, Beijing, Beijing, China •Peking University Sixth Hospital, Beijing, Beijing, China •The Second Xiangya Hospital of Central South University, Changsha, Hunan, China •Shanghai Mental Health Center, Shanghai, Shanghai, China •Xi'an Mental Health Center, Xi'an, Shanxi, China •West China Hospital, Sichuan Univeristy, Chengdu, Sichuan, China •Tianjin Mental Health Center, Tianjin, Tianjin, China •The First Affiliated hospital of Zhejiang University School of Medicine, Hangzhou, Zhejiang, China

PAPER A

Title

Perspectives on machine learning for classification of Schizotypy using fMRI data

Authors

Madsen, Kristoffer H; Krohne, Laerke G ; Cai, Xin-Lu; Wang, Yi; Chan, Raymond CK;

Journal

Schizophrenia Bulletin

Year

2018

Perspectives on Machine Learning for Classification of Schizotypy Using fMRI Data

Kristoffer H. Madsen[*,1,2], **Laerke G. Krohne**[1,2], **Xin-lu Cai**[3-5], **Yi Wang**[3], and **Raymond C. K. Chan**[3-6]

[1]Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Hvidovre, Denmark; [2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark; [3]Neuropsychology and Applied Cognitive Neuroscience Laboratory, CAS Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China; [4]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China; [5]Sino-Danish College, University of Chinese Academy of Sciences, Beijing, China; [6]These authors shared correspondence to this work.

*To whom correspondence should be addressed; tel: +45 38622975; fax:+45 36351680; e-mail: kristofferm@drcmr.dk.

Functional magnetic resonance imaging is capable of estimating functional activation and connectivity in the human brain, and lately there has been increased interest in the use of these functional modalities combined with machine learning for identification of psychiatric traits. While these methods bear great potential for early diagnosis and better understanding of disease processes, there are wide ranges of processing choices and pitfalls that may severely hamper interpretation and generalization performance unless carefully considered. In this perspective article, we aim to motivate the use of machine learning schizotypy research. To this end, we describe common data processing steps while commenting on best practices and procedures. First, we introduce the important role of schizotypy to motivate the importance of reliable classification, and summarize existing machine learning literature on schizotypy. Then, we describe procedures for extraction of features based on fMRI data, including statistical parametric mapping, parcellation, complex network analysis, and decomposition methods, as well as classification with a special focus on support vector classification and deep learning. We provide more detailed descriptions and software as supplementary material. Finally, we present current challenges in machine learning for classification of schizotypy and comment on future trends and perspectives.

Key words: functional magnetic resonance imaging/feature extraction/neuroimaging/schizotypy/schizophrenia spectrum disorder

Introduction

The study of schizotypy has received substantial interest in the field of psychiatry and psychology and recently developments and increased interest in machine learning for neuroimaging is showing promising applications in computational psychiatry. Theoretically, schizotypy has been conceptualized as an important phenotype for schizophrenia spectrum disorders.[1–3] Two competitive theories, the quasi-dimensional and the fully dimensional approach have been proposed to model the construct of schizotypy. The quasi-dimensional approach posits the view that schizotypy is a discontinuity in the general population.[2,4] However, recent studies have suggested that this phenotype is distributed along a continuum, ranging from psychological well-being to full-blown psychosis,[5–7] supporting the fully dimensional approach that emphasizes the continuity of schizotypy.[8] Furthermore, empirical findings demonstrate that individuals with schizotypal traits exhibit similar but attenuated impairments in cognition,[9,10] emotion,[7,11] and neurological functions[10,12] compared with patients with schizophrenia. Likewise, manifestations of these schizotypal phenotypes are found to be robust and stable across time and environment.[5,13–15]

With implications from the neurodevelopmental model of psychosis in schizophrenia,[16,17] Insel[18] further delineated 4 stages, ranging from risk to chronic disability. This 4-stage hypothesis highlights the importance of early risk stages for the understanding of the psychopathology to facilitate early detection and intervention strategies for psychosis and mental disorders. Although schizotypy is not explicitly included in Insel's model, there are important similarities within the cognitive, emotional, and social impairments. This point toward understanding the psychopathology of schizophrenia spectrum

disorders through personality traits presented in the general population.

Recently, schizotypy has recently been conceptualized as a construct well beyond the borders of schizophrenia spectrum disorders (e.g., Cohen et al[19]). These authors argue that the researchers interested in psychosis have mainly followed narrow research avenues, focusing on molecular, neurophysiological, environmental, and cultural correlates of psychotic expression or investigating potential endophenotypes relating to the extreme manifestation of schizotypy to schizophrenia. However, the unique emotional and social manifestations observed in individuals with schizotypy can actually provide insight into the nature of affective and social systems integral to general human functioning. For example, findings from functional neuroimaging have shown that individuals with social anhedonia exhibit significant hypoactivation of the left pulvinar, claustrum, and insula to positive cues in the anticipatory phase of the affective incentive delay task compared with those without social anhedonia.[14] Longitudinal studies also suggest that individuals with schizotypal traits have their unique trajectories that may not necessarily develop into full-blown psychosis.[20–23] Recently, Wang et al[24] identified 4 trajectories of schizotypy; including 2 stable and 2 reactive groups. The "stable low and high schizotypy" groups displayed the best and worst clinical and functional outcomes, respectively. The "high reactive schizotypy" group was characterized by a relatively rapid decline in function, while the "low reactive schizotypy" group was characterized by low scores at baseline of the assessment but with gradual deterioration. These findings suggest that even within the nonclinical sample of schizotypal phenotypes, similar subtypes, and trajectories comparable to the clinical patients with schizophrenia are observed. This highlights the importance of tracking schizotypy longitudinally because of their unique trajectories and outcomes.

Several studies have already applied neuroimaging data to investigate the neurobiological changes related to schizotypy, reporting both structural and functional changes. For example, structural studies have found grey matter volume changes in many of the areas known to be altered in schizophrenia, such as prefrontal, temporal, and cingulate cortex, as well as insula and subcortical regions.[25–28] These studies suggest that cortical changes exist on a dimensional continuum across the schizophrenia spectrum, likely to occur pre-onset of psychopathology. Furthermore, studies using functional magnetic resonance imaging (fMRI) to investigate social cognition, have reported similar regional brain activation changes, when comparing participants with different degree of schizotypy, or individuals with high schizotypy compared with controls.[29–31] Finally, functional connectivity studies reported similar network changes to that of patients with schizophrenia,[32–34] such as altered connectivity between striatum, medial prefrontal cortex (PFC), anterior cingulate (ACC), and insula. Importantly, almost all the above studies, reported different results for

the positive and negative dimension of schizotypy, demonstrating the heterogeneity of schizotypy.

The above findings emphasize the important role of schizotypy in psychiatry and psychology. On one hand, schizotypy is considered to be a trait marker for schizophrenia and the study of behavioral and neurobiological bases of schizotypy may help us understand the underlying psychopathology of schizophrenia. This suggests that schizotypy may be an important phenotype for studying schizophrenia spectrum disorders. On the other hand, schizotypy may serve as a unique entity to examine the underlying emotional and social systems in humans. Therefore, a better way to classify this phenotype will be meaningful to schizotypy scholars. However, to our knowledge, there are only few studies which have been identifying schizotypy based on neuroimaging data. Here, machine learning methods can serve to bridge this knowledge gap, and help elucidate the neurobiological abnormalities of at-risk individuals at an early stage of schizophrenia.

Machine Learning in the Field of Schizotypy and Schizophrenia

The overall aim of machine learning is to make computers classify data without being explicitly programmed. Typically, a distinction is made between supervised and unsupervised learning. The former refers to learning using labelled data, with the aim to generalize classification to data with unknown labels. In contrast, unsupervised learning methods explore statistical dependencies in unlabelled data, with the goal of learning structure in the data and possibly cluster data into distinct classes.

Recently, machine learning methods have been used as a neuroimaging-based tool to automatically discriminate individuals in schizophrenia spectrum disorders from healthy people.[35–37] Empirical findings suggest that these methods are able to classify schizophrenia patients from healthy controls with an accuracy rate ranging from 75% to 98%.[37–40] Furthermore, recent studies have had success with using support vector classification (SVC) to predict the transition of ultra-high risk individuals converting to full-blown psychosis,[41–43] and discriminate converters and nonconverters.[44,45] However, limited studies have investigated individuals in the stages before onset of the illness.

As for studying schizotypy using machine learning methods, a range of studies have been exploring the neural mechanism related to schizotypy and classified individuals according to different groups. In 2006, Shinkareva et al[46] used spatio-temporal dissimilarity maps to classify individuals with high levels of positive schizotypy and controls based on fMRI data from an emotional Stroop task. With the same aim, Modinos et al[47] performed SVC on brain activation maps from an emotional task and found the alterations for the emotional circuitry, including amygdala, ACC, and medial PFC, in individuals with

high positive schizotypy. For comparison, they also performed statistical parametric mapping (SPM), which did not detect any class differences, indicating the increased sensitivity to subtle changes in risk populations, by using multivariate approaches. From the view of the "full dimensional" model of schizotypy, Wiebels et al[48] used partial least square method, to demonstrate the relationship between different facets of schizotypy with gray matter volume changes.

Furthermore, 2 studies have explored schizotypal scores in individuals with subclinical depression and an ultra-high-risk group, respectively. First, Modinos et al[49] found significant correlation between the positive dimension of schizotypy and the SVC weights which were obtained when classifying individuals with subclinical depressive symptoms and healthy controls. Secondly, in a longitudinal study, Zarogianni et al[45] applied SVC to classify ultra-high-risk individuals into converters and nonconverters. Whereas this study mainly used structural MRI data, it was shown that the classification performance was increased when adding schizotypy scores to the analysis. Finally, other neuroimaging modalities than (f)MRI have been used to investigate schizotypy using machine learning methods. For example, in a study by Jeong et al[50] event-related potentials, measured by EEG during an audiovisual emotion perception task, were used for classification of individuals with schizotypy and controls.

To conclude, the research in schizotypy utilizing machine learning shows great promise in terms of improving our understanding of schizotypy, and is of particular relevance for early detection and potential interventions. The main advantage of machine learning methods is that they can offer higher sensitivity than their counterparts based on standard univariate statistics, due to being able to learn the likely complex manifestations of schizotypy in multimodal neuroimaging data. Currently, existing studies are still limited by quite small sample sizes ($n = 7–18$ in each group[45–47,49,50]), and there is a risk that the reported classification rates are overfitted to the observed samples. This highlights the importance of having sufficient large sample sizes, and well-balanced groups to enable adequate learning and ensure that the training data is representative. Furthermore, it is important that future studies focus on independent validation of existing results to ensure that findings are generalizable to the population.

Classification and Feature Extraction Methods

In neuroimaging studies, fMRI data are mostly used to measure either activation changes in isolated brain areas, or to estimate functional connectivity (networks coupling) across regions.[51] Because fMRI data are recorded in relatively high spatial resolution with a limited number of time points, estimation of activation patterns and in particular connectivity, is in practice quite unstable.[52] Therefore, approaches to reduce dimensionality are often considered to improve the stability of the estimated functional activation.[53–55] In the current article, we focus on features derived from fMRI, but classification procedures readily generalize to other modalities and multi-modal settings.

When using supervised learning in the field of neuroimaging, the aim is generally to determine an unknown class label of a subject based solely on the measured imaging data (eg, recorded fMRI data) as illustrated in
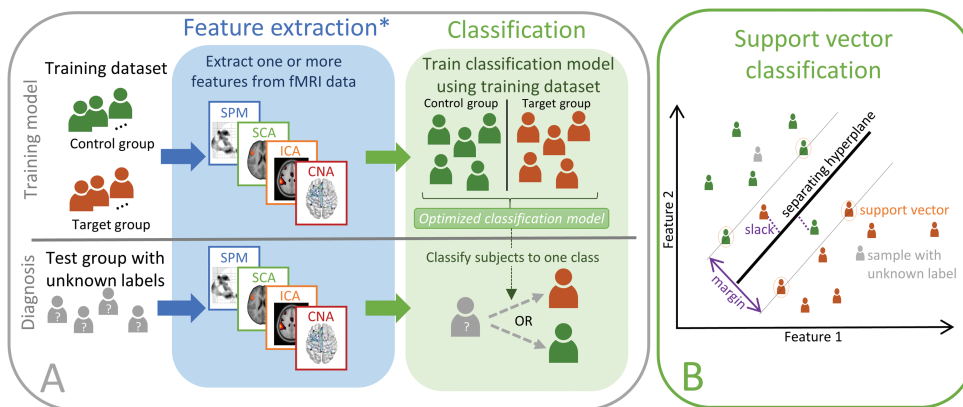
Fig. 1. Classification. The top row in panel A shows how a classification model can be trained on neuroimaging data. First feature extraction methods are used to identify features that can be used to train a classification model on samples with known labels. Once the classification model is trained, it can be applied to features extracted (using the same procedure) from subjects with unknown labels as indicated in the bottom row. *In principle the feature extraction step can be omitted. However, in practice for many imaging modalities (including fMRI), overfitting due to the high dimensionality of the input data will be detrimental to the classification performance. Panel B provides an illustration of the linear soft margin SVC algorithm in a 2-dimensional feature space. The SVC identifies the separating hyperplane that maximizes the margin, this hyperplane is only defined by the support vectors which are samples that are on the margin (marked by a circle). The soft margin SVC allows misclassification to avoid overfitting by introducing slack variables for each misclassified sample (marked with a dotted line). When the SVC is trained the labels of new samples (marked in gray) can be estimated according to the side of the hyperplane on which they reside.

figure 1, this procedure is also termed classification. In supervised classification, a model discriminating between the known labels in the training data is learned, subsequently enabling application of this model to unlabelled data to predict unknown labels.

Given a labelled dataset, one can determine the classification performance using cross-validation (CV). The accuracy (rate of correctly identified class labels) is often used as a measure of performance. However, is important to note that this does not provide a full description of the performance, but also sensitivity (also referred to as the true positive rate or recall) and the specificity (the true negative rate) are important quantities. To test if the obtained classification rate is significant, the performance is usually tested against a parametric or empirical null-distribution.[56] If the classification step considers several separate classification procedures, corrections for multiple comparisons should be performed when assessing significance. The CV procedure can be considered a simulation of the clinical setting, in which the labels of a set of subjects (test set) are assumed unknown and to be estimated through the training of a classification algorithm on the remaining subjects (training set). A frequently used method is the leave-one-out CV, where only one subject constitutes the test set, and the procedure is repeated for each subject as illustrated in figure 3. The leave-one-out scheme is often preferred because it minimizes the model bias by reserving the maximal amount of data for model training, but is has the disadvantage that there is a higher risk of overfitting to the training data. Therefore, other schemes such as *K*-fold (dividing the data

into *K* nonoverlapping splits) CV are sometimes preferred. These enable testing of model stability by examining the variability of the identified model across splits. An example is the split half resampling procedure, where the difference between the models in the 2 independent splits can serve as an estimate of the model reproducibility.[57]

In principle, it is possible to train classification algorithms directly on the raw neuroimaging data. However, due to the high dimensionality of the data compared with the small sample size the input data will appear sparse in the high dimensional space, often referred to as the curse-of-dimensionality. This in turn causes the classification procedure to be too specialized and generalize poorly to the test data, a phenomenon known as overfitting.

Therefore, classification is typically approached using a 2-step procedure in which features relevant to classification are first identified (see feature extraction steps, as illustrated in figure 2) and subsequently used to train the classification algorithm. The feature extraction step might include feature selection, where features are selected for further training. It is important that feature selection should only use labels from the training dataset, as the evaluation of performance would otherwise be biased and potentially lead to overfitting. Therefore, nested CV schemes, where an additional independent test set is used to estimate optimal features or other free parameters can be advantageous. Overfitting may be mitigated by automated feature selection methods and ensemble learning methods[58] such as forward selection, backward elimination, recursive feature elimination,[59] decision trees, and
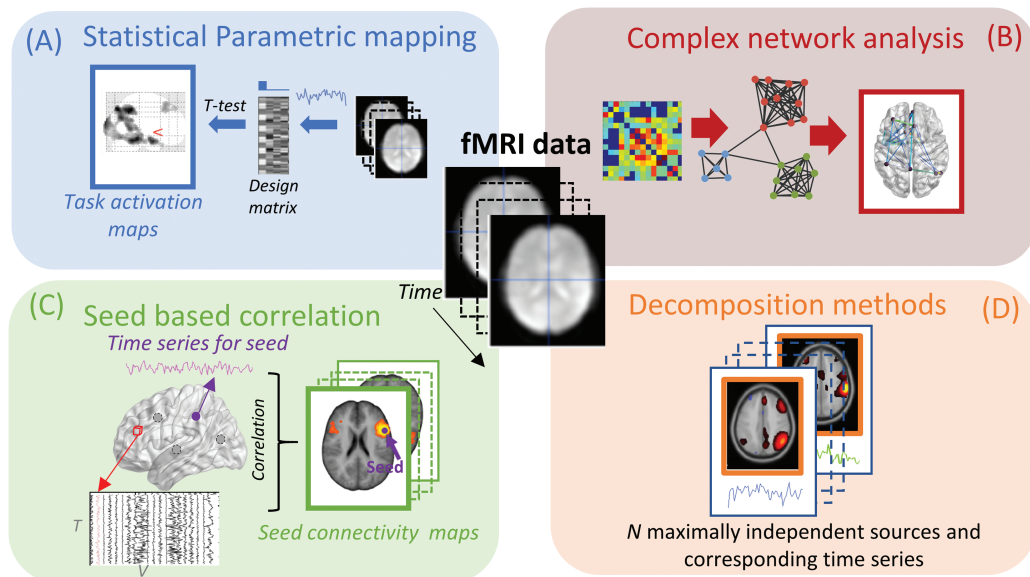
Fig. 2. Sketch of 4 feature extraction methods for fMRI. Panel A illustrates statistical parametric mapping, where information about the experimental design is used to test for significant activation in each voxel using a general linear model. Panel B sketches complex network analysis. Here, a network is derived by determining functional connectivity between parcellated brain regions, followed by analysis using graph theoretical measures. In panel C, the seed based correlation approach is illustrated, here the time series from a predetermined brain region is extracted and correlated to the rest of the brain. In panel D, decomposition methods are illustrated where fMRI data are decomposed into spatially independent components with corresponding time series.

Leave-one-out cross validation
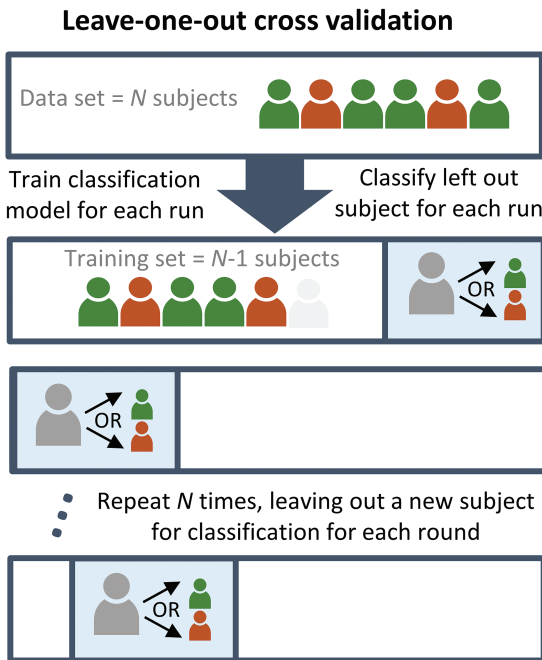
Fig. 3. Leave-one-out cross-validation. The figure illustrates the leave-one-out LOO CV procedure. For each participant a classification model excluding that particular participant is trained. The model is then used to estimate the class label of the participant. This procedure is repeated for each participant to provide an unbiased estimate of the classification performance. Note that other CV schemes, including more complex nested CV are also possible.

random decision forests.[60] Also, several toolboxes including scikit-learn,[61] Nilearn,[62] PRoNTo,[63] pyMVPA,[64] and the NeuroMiner toolbox used by Koutsouleris et al[41] are tailored toward machine learning for neuroimaging and provide tools for automated feature selection.

Appropriate preprocessing steps are very important before feature extraction, since data which are contaminated by artefacts might not only lead to poor classification performance, but may also cause difficulties in the interpretation of the results. For example, if movement artefacts are more dominant in one of the groups, the classifier might focus on movement artefacts and obtain good classification performance. For further information on common preprocessing steps and software, see supplementary section A.

In the following subsections, we describe a selection of often used feature extraction procedures, and although not covered by this article, additional methods exist, including fALFF,[65,66] and methods for estimating regional signal homogeniety.[67]

Statistical Parametric Mapping

SPM is currently one of the most frequent used methods for analyzing task-based fMRI data. The overall goal of SPM is to localize brain activation that differs significantly

between tasks[68] as illustrated in figure 2A. The technique is mass univariate, which means that an independent parametric statistical test (t- or f-test) is performed for each voxel separately, typically using the general linear model. The 3 most common software packages for performing parametric mapping are SPM,[69] FSL,[70] and AFNI.[71]

When used for classification, the parameter estimates or statistical values (extracted across the entire brain or in regions of interest) are used as classification features, either directly or with an additional feature selection step. An advantage of using SPMs is that the localization of effects is already implied in the features, typically leading to more straightforward interpretation of models. However, because the procedure is essentially univariate it can miss important information shared across a range of variables, and therefore may be less sensitive than feature extraction methods that consider the multivariate structure of the data directly.

Parcellation, Complex Networks, and Seed-Based Analysis

To overcome instability problems due to the low temporal resolution as described above, approaches that parcellate the brain into fewer regions; either defined via atlases[72,73] or from data driven clustering methods[53,74,75] are often preferred. Functional connectivity features can then be determined between the parcels using a statistical measure such as (partial) correlation or mutual information. The resulting features (typically represented in a symmetric adjacency matrix representing the network coupling between each parcel) can either serve directly as feature for subsequent classification or be used for further extraction of features, eg, in a graph theoretical framework (figure 2B). Often the graphs are binarized by applying a threshold, and global measures such as the node degree distribution (number of connections between parcels/nodes) graph structure via modularity[76] or relational modeling[77,78] are used to characterize networks. For a more complete description of graph theoretical measures see Bullmore and Sporns.[79]

A related technique is the simple and intuitive seed-based correlation analysis (SCA),[51] which determines the coupling between a number of predefined seeds (based on some a priori hypothesis, from either a localization experiment or the literature). The time series data from each seed is then correlated with all other voxels of the brain, resulting in a whole-brain, voxel-wise functional connectivity map for each seed as shown in figure 2C. For a more detailed description of SCA and how it has been used in resting state fMRI in comparison with data driven methods, please see Cole et al.[80] In general, parcellation-based methods are attractive because they generate a more simplistic overview of the data, and often lead to more straightforward interpretation of features. However, the limited flexibility that is implied by fixed parcellation

schemes can lead to selection of inappropriate features and result in decreased sensitivity.

Decomposition

Decomposition are unsupervised machine learning methods (sometimes also referred to as data driven methods) that seek to identify latent sources in the data from multiple measurements (ie, fMRI time series). In fMRI, this typically amounts to identifying a relatively low number of underlying spatial sources (typically between 10 and 100) that are associated with time series as illustrated in figure 2D. The procedure can be viewed as a (lossy) compression of the information in the data. The sources are often considered representations of functional networks, because they represent consistent time courses across the brain. A widely accepted method is spatial group independent component analysis (ICA), which results in individual component expressions (sources) across subjects with corresponding time series. Most frequently ICA is performed using one of the open source toolboxes, such as GIFT[81] or FSL Melodic.[82] Decomposition is advantageous because consistent activation patterns can be captured efficiently and automatically. A potential disadvantage is that interpretation can be challenging because decomposition is prone to also capture prominent nuisance effects in the data including motion and physiological signals such as the cardiac and respiratory cycles. Also, there are typically a wide range of adjustable parameters (such as the number of sources) that are difficult to set manually and can lead to overfitting if considered part of the learning algorithm. More information on decomposition is provided in supplementary section B.

Support Vector Classification

Supervised classification methods seek to identify some function that would enable discrimination between the labels in the training dataset. Importantly, when the input dimensionality is high compared to the number of samples (typically the case in fMRI unless elaborate feature extraction and selection has taken place), it is actually trivial to obtain perfect classification in the training set (overfitting), but the performance may generalize poorly to the test set. Therefore, the real challenge in classification is to ensure that the classification generalizes well to unseen samples.

There are many classification algorithms available. Here, we will focus on the SVC methods,[83,84] because they have often been used in previous literature and are readily available in several easy to use software packages.[61,85] For further reference and information on other classification methods, see, eg, Schmah et al.[86] and Bishop.[87]

The simplest classification problem is a binary (2 classes) linear classification, where the SVC algorithm attempts to identify a discrimination function expressed as a linear projection across features, where the sign indicates the label. This is most straightforwardly illustrated in the

2-dimensional case, where the so-called separating hyperplane is a straight line (figure 1B), here it is also evident that there are many lines that would lead to identical classification performance. The SVC chooses the hyperplane that maximizes the margin, ie, the perpendicular distance between the plane the closest data points. The SVC therefore focuses on the points on the margin (samples that are the hardest to classify, also called support vectors), and the classification of new samples only require information about the distances with respect to these so-called support vectors, allowing efficient evaluation. This is often referred to as a solution with sparse support in the training set, where sparsity here refers to samples rather than features. In practice, the soft margin SVC[83] is mostly preferred as it allows misclassified samples, to obtain a larger margin, which will increase the stability of the classifier. In this case, the maximization of the margin is traded off against a penalty for misclassified samples which is proportional to the distance to the separating hyperplane. The trade-off is controlled by a parameter (typically referred to as the C-parameter), which has to be selected or determined through an additional nested CV procedure.[85] For unbalanced dataset (cases in which the no. samples in each group differs) the class imbalance can be taken into account by weighting the hyperplane such that the imbalance is counteracted (by assigning more weight to the under-represented class). Also, for such datasets the accuracy alone may not be a good performance measure, as even a trivial classifier that always selects the most frequent label would appear to perform well. In these cases, using other metrics such as prediction-recall curves or Matthew's correlation coefficient are usually more informative.[88]

Generalization to nonlinear discrimination is typically approached by projecting the data into another space (higher dimensional or even infinitely dimensional) which would enable linear separation. For the SVC, and a range of other classification methods, this can be efficiently implemented using kernels through the so-called kernel-trick. Here, it is sufficient to calculate distances between the samples measured in the projected space (represented in a Gram matrix) which circumvents operating with the potentially high dimensional projection explicitly. Examples of frequently used kernels are the linear kernel (for linear classification), radial basis function kernel and the polynomial kernel. It is important to note that kernels typically introduce at additional parameters that need to be selected or optimized via CV,[85] which can exacerbate problems of overfitting.

The classification performance is rarely the only quantity of interest. Often researchers are interested in determining which brain regions are important for classification. For the linear SVC weight maps, or sensitivity maps[89] for nonlinear classifiers, are often visualized, as they indicate the importance of each feature for the classification performance. The interpretation of these weight maps is not straightforward as features can actually be

important for classification, not because they are directly related to the effect of interest, but rather because they serve to filter out nuisance effects. This issue was highlighted by Haufe et al[90] suggesting a procedure for transforming weight maps into more interpretable visualizations for linear classification.

In practice, data labels (eg, schizotypy score) are often determined using questionnaires, which utilizes either continuous or ordinal scales, where it might be difficult to define a clear division between classes. In such cases, it can be attractive to train the algorithms to predict this continuous variable directly. This effectively turns the procedure into a multivariate regression problem. Here, support vector regression[91] is analogous with SVC, where the margin is formed by considering how far the predicted value (in the training set) is from the measured value. When considering regression models in place of classification, it should be noted that other performance measures such as the mean absolute error have to be used. Unfortunately, the interpretation of such measures is in general less intuitive than classification rates. Furthermore, evaluation of statistical significance is more involved and researcher most often rely on random permutation tests to form empirical null-distributions.[56]

To illustrate the classification procedures described above, we have added an illustrative example to the supplementary material, where we used SVC to classify participants into either a low of high social anhedonia group, using features from both SPM and ICA. For more details, please refer to supplementary material section A.

Deep Learning

Deep learning based on neural networks have recently received much attention in the machine learning community, and have also been used to classify neuroimaging data in several general[92,93] and clinical settings.[94–96] The general philosophy behind deep learning is to train large neural networks with many layers and parameters that take the raw (or in most cases preprocessed) data as input and where the last layer in the network produces an outcome such as classification of subjects. If properly trained the first layers of the network should then represent basic features of the data, that are then refined and specialized in the subsequent layers. As these networks inherently contain many parameters overfitting due to the limited amount of data is a major concern when attempting to train networks. Here, mitigation strategies include regularization, dropout sampling, and weight sharing.[97] Another option, is to use transfer learning approaches, which use networks that are pretrained on other datasets (which may even be of a different modality) and only refine weights in the last layers of the network.[98] We believe that such strategies, potentially combined with data argumentation[99] (where more samples are created using transformations/perturbations of the original data), will

be extremely important in the future to ensure the success of deep learning in schizotypy research.

Discussion and Future Perspectives

In the preceding paragraphs, we have motivated the importance for classification of schizotypy, presented previous literature that has used machine learning methods for classification, and described methods for feature extraction and classification. Machine learning approaches have a range of advantages, which make them very attractive for studying early risk stages and subtle differences, as it is the case for schizotypy. A clear example of how these methods can increase the sensitivity to subtle changes, was shown in Modinos et al,[47] who found significant alterations in an emotional circuitry in individuals with schizotypy when using SVC, whereas no class difference was detected when using a standard SPM analysis.

However, even though machine learning methods have shown very promising results so far, there are a wide range of pitfalls and challenges that needs to be considered. In the following, we will highlight some of the most important aspects, which should be kept in mind when using machine learning methods for classification of schizotypy or similar early risk groups.

The high dimensionality and typical low sample sizes available in studies represent a challenging problem for machine learning algorithms. Thus, procedures to reduce dimensionality of the input data (feature extraction) and regularization are necessary to ensure good generalization performance. While repeated nested CV procedures are useful for tentatively alleviating data availability issues, initiatives to encourage data sharing across sites are very important to overcome the problems of sparse sample availability.[100,101]

Appropriate pre-processing can have a profound impact on results and a wide range of choices are available both in terms of methods and parameters.[102] This is also true for feature extraction, feature selection and classification steps, and it is important to note that if these choices are considered free parameters of the classification, the problem of overfitting is exacerbated and appropriate procedures to improve generalization such as CV should be considered. The feature extraction method of choice will depend on the research question. If the study is driven by specific hypotheses, it can be an advantage to use feature extraction methods that specifically extract the relevant dimensions of the data. Whereas, if the study is more explorative, decomposition methods may be preferred, as it avoids restriction of the analysis to a set of predefined hypotheses.

In general, the high degree of flexibility in choices of classification pipelines represents a challenge. It is very difficult for researcher to prove that none of the choices biased the reported classification performance (because the pipeline was optimized for classification performance), which might happen even inadvertently. To

circumvent these issues, it is highly recommendable that specific hypotheses and detailed analysis procedures are preregistered before studies are commenced. This can be done easily using, eg, the Open Science Framework (https://osf.io/). Note that such preregistration is valuable even for studies with explorative hypotheses. In addition, it is obviously important that studies with negative outcomes are also published, and that specific studies that seek to reproduce previous findings are commenced.

Schizophrenia spectrum disorders are complex and consist of a wide range of symptoms with heterogeneous disease progression across individuals. In practice, this poses challenges in clearly defining disease phenotypes and renders interpretation of potential results difficult. The view of schizotypy as a continuous range of symptoms and traits expressed by individuals, motivate the use of machine learning to predict multiple continuous measures of disease progression. Here, it is natural to consider multivariate regression models, such as support vector regression,[91] to directly predict schizotypy traits. Also, to take advantage of the fact that multiple dimensions of schizotypy are often assessed using a variety of rating scales, methods such as partial least squares regression[103] can be used to establish compact relations between multivariate neuroimaging data and multiple schizotypy measures.

The use of these tools and more generally applicable methods based on deep learning, represent promising research avenues, which can help us gain a more complete understanding of schizotypy, lead to improved identification of individuals with schizotypy and facilitate appropriate management and intervention for these individuals. Machine learning constitutes a paradigm shift toward quantitative evaluation, where we no longer need to rely on subjective rating and structured interviews. Consequently, the time spend on identification of subtypes of schizophrenia spectrum disorders can be reduced while potentially improving the accuracy in clinical practice.

In summary, classification of schizotypy represents a promising application for the combination of machine learning and neuroimaging, but there are still a range of challenges, in particular, related to how robustness to overfitting, and thereby better generalization performance can be archived. However, if these challenges are appropriately addressed, machine learning can significantly improve our understanding of schizotypy and schizophrenia spectrum disorders. Finally, the emerging field of computational psychiatry had important applications in disease prevention, early diagnosis, identification of drug targets, and individual treatment plans for psychiatric diseases and may revolutionize modern neurology.

Supplementary Material

Supplementary data are available at *Schizophrenia Bulletin* online.

References

1. Meehl PE. Schizotaxia, schizotypy, schizophrenia. *Am Psychol*. 1962;17:827–838.
2. Meehl PE. Toward an integrated theory of schizotaxia, schizotypy, and schizophrenia. *J Pers Disord*. 1990;4:1–99.
3. Debbané M, Barrantes-Vidal N. Schizotypy from a developmental perspective. *Schizophr Bull*. 2015;41:S386–S395.
4. Everett KV, Linscott RJ. Dimensionality vs taxonicity of schizotypy: some new data and challenges ahead. *Schizophr Bull*. 2015;41(suppl 2):S465–S474.
5. Kwapil TR, Barrantes-Vidal N, Silvia PJ. The dimensional structure of the Wisconsin Schizotypy Scales: factor identification and construct validity. *Schizophr Bull*. 2008;34: 444–457.
6. Chan RCK, Li X, Lai M, et al. Exploratory study on the base-rate of paranoid ideation in a non-clinical Chinese sample. *Psychiatry Res*. 2011;185:254–260.
7. Wang Y, Lui SSY, Zou LQ, et al. Individuals with psychometric schizotypy show similar social but not physical anhedonia to patients with schizophrenia. *Psychiatry Res*. 2014;216:161–167.
8. Claridge G, Beech T. Fully and quasi-dimensional constructions of schizotypy. In: *Schizotypal Personality*. New York, NY: Cambridge University Press; 1995:192–216. http://psyc-net.apa.org/doi/10.1017/CBO9780511759031.010
9. Wang Y, Chan RC, Xin Yu, Shi C, Cui J, Deng Y. Prospective memory deficits in subjects with schizophrenia spectrum disorders: a comparison study with schizophrenic subjects, psychometrically defined schizotypal subjects, and healthy controls. *Schizophr Res*. 2008;106:70–80.
10. Ettinger U, Mohr C, Gooding DC, et al. Cognition and brain function in schizotypy: a selective review. *Schizophr Bull*. 2015;41(suppl 2):S417–S426.
11. Lui SSY, Liu ACY, Chui WWH, et al. The nature of anhedonia and avolition in patients with first-episode schizophrenia. *Psychol Med*. 2016;46:437–447.
12. Chan RCK, Cui H, Chu M, et al. Neurological soft signs precede the onset of schizophrenia: a study of individuals with schizotypy, ultra-high-risk individuals, and first-onset schizophrenia. *Eur Arch Psychiatry Clin Neurosci*. 2017;268:49–56.
13. Chan RCK, Song Shi H, Lei Geng F, et al. The Chapman psychosis-proneness scales: consistency across culture and time. *Psychiatry Res*. 2015;228:143–149.
14. Chan RCK, Li Z, Li K, et al. Distinct processing of social and monetary rewards in late adolescents with trait anhedonia. *Neuropsychology*. 2016;30:274–280.
15. Fonseca-Pedrero E, Debbané M, Ortuño-Sierra J, et al. The structure of schizotypal personality traits: a cross-national study. *Psychol Med*. 2018;48:451–462.
16. Murray RM, Lewis SW. Is schizophrenia a neurodevelopmental disorder? *Br Med J (Clin Res Ed)*. 1987;295:681–682.

17. Murray RM, Lewis SW. Is schizophrenia a neurodevelopmental disorder? *Br Med J (Clin Res Ed)*. 1988;296:63.

18. Insel TR. Rethinking schizophrenia. *Nature*. 2010;468:187–193.

19. Cohen AS, Mohr C, Ettinger U, Chan RC, Park S. Schizotypy as an organizing framework for social and affective sciences. *Schizophr Bull*. 2015;41(suppl 2):S427–S435.

20. Gooding DC, Tallent KA, Matts CW. Clinical status of at-risk individuals 5 years later: further validation of the psychometric high-risk strategy. *J Abnorm Psychol*. 2005;114:170–175.

21. Gooding DC, Tallent KA, Matts CW. Rates of avoidant, schizotypal, schizoid and paranoid personality disorders in psychometric high-risk groups at 5-year follow-up. *Schizophr Res*. 2007;94:373–374.

22. Kwapil TR. Social anhedonia as a predictor of the development of schizophrenia-spectrum disorders. *J Abnorm Psychol*. 1998;107:558–565.

23. Kwapil TR, Gross GM, Silvia PJ, Barrantes-Vidal N. Prediction of psychopathology and functional impairment by positive and negative schizotypy in the Chapmans' ten-year longitudinal study. *J Abnorm Psychol*. 2013;122:807–815.

24. Wang Y, Shi H, Liu W, et al. Trajectories of schizotypy and their emotional and social functioning: an 18-month follow-up study. *Schizophr Res*. 2017. doi:10.1016/j.schres.2017.07.038.

25. Ettinger U, Williams SC, Meisenzahl EM, Möller HJ, Kumari V, Koutsouleris N. Association between brain structure and psychometric schizotypy in healthy individuals. *World J Biol Psychiatry*. 2012;13:544–549.

26. Wang Y, Yan C, Yin DZ, et al. Neurobiological changes of schizotypy: evidence from both volume-based morphometric analysis and resting-state functional connectivity. *Schizophr Bull*. 2015;41:S444–S454.

27. Modinos G, Mechelli A, Ormel J, Groenewold NA, Aleman A, McGuire PK. Schizotypy and brain structure: a voxel-based morphometry study. *Psychol Med*. 2010;40:1423–1431.

28. Modenato C, Draganski B. The concept of schizotypy—a computational anatomy perspective. *Schizophr Res Cogn*. 2015;2:89–92.

29. Modinos G, Ormel J, Aleman A. Altered activation and functional connectivity of neural systems supporting cognitive control of emotion in psychosis proneness. *Schizophr Res*. 2010;118:88–97.

30. Wang Y, Liu W, Li Z, et al. Dimensional schizotypy and social cognition: an fMRI imaging study. *Front Behav Neurosci*. 2015;9:133.

31. Kanske P, Böckler A, Trautwein FM, Parianen Lesemann FH, Singer T. Are strong empathizers better mentalizers? Evidence for independence and interaction between the routes of social cognition. *Soc Cogn Affect Neurosci*. 2016;11:1383–1392.

32. Wang Y, Liu WHH, Li Z, et al. Altered corticostriatal functional connectivity in individuals with high social anhedonia. *Psychol Med*. 2016;46:125–135.

33. Wang Y, Ettinger U, Meindl T, Chan RCK. Association of schizotypy with striatocortical functional connectivity and its asymmetry in healthy adults. *Hum Brain Mapp*. 2018;39:288–299.

34. Lagioia A, Van De Ville D, Debbané M, Lazeyras F, Eliez S. Adolescent resting state networks and their associations with schizotypal trait expression. *Front Syst Neurosci*. 2010;4:1–12.

35. Ardekani BA, Tabesh A, Sevy S, Robinson DG, Bilder RM, Szeszko PR. Diffusion tensor imaging reliably differentiates patients with schizophrenia from healthy volunteers. *Hum Brain Mapp*. 2011;32:1–9.

36. Janousova E, Montana G, Kasparek T, Schwarz D. Supervised, multivariate, whole-brain reduction did not help to achieve high classification performance in schizophrenia research. *Front Neurosci*. 2016;10:392.

37. Kaufmann T, Skåtun KC, Alnæs D, et al. Disintegration of sensorimotor brain networks in schizophrenia. *Schizophr Bull*. 2015;41:1326–1335.

38. Arbabshirani MR, Kiehl KA, Pearlson GD, Calhoun VD. Classification of schizophrenia patients based on resting-state functional network connectivity. *Front Neurosci*. 2013;7:133.

39. Chyzhyk D, Savio A, Graña M. Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of ELM. *Neural Netw*. 2015;68:23–33.

40. Du W, Calhoun VD, Li H, et al. High classification accuracy for schizophrenia with rest and task fMRI data. *Front Hum Neurosci*. 2012;6:145.

41. Koutsouleris N, Meisenzahl EM, Davatzikos C, et al. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Arch Gen Psychiatry*. 2009;66:700–712.

42. Koutsouleris N, Borgwardt S, Meisenzahl EM, Bottlender R, Möller HJ, Riecher-Rössler A. Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study. *Schizophr Bull*. 2012;38:1234–1246.

43. Nejad AB, Madsen KH, Ebdrup BH, et al. Neural markers of negative symptom outcomes in distributed working memory brain activity of antipsychotic-naive schizophrenia patients. *Int J Neuropsychopharmacol*. 2013;16:1195–1204.

44. Koutsouleris N, Davatzikos C, Bottlender R, et al. Early recognition and disease prediction in the at-risk mental states for psychosis using neurocognitive pattern classification. *Schizophr Bull*. 2012;38:1200–1215.

45. Zarogianni E, Storkey AJ, Johnstone EC, Owens DG, Lawrie SM. Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. *Schizophr Res*. 2017;181:6–12.

46. Shinkareva SV, Ombao HC, Sutton BP, Mohanty A, Miller GA. Classification of functional brain images with a spatio-temporal dissimilarity map. *Neuroimage*. 2006;33:63–71.

47. Modinos G, Pettersson-Yeo W, Allen P, McGuire PK, Aleman A, Mechelli A. Multivariate pattern classification reveals differential brain activation during emotional processing in individuals with psychosis proneness. *Neuroimage*. 2012;59:3033–3041.

48. Wiebels K, Waldie KE, Roberts RP, Park HR. Identifying grey matter changes in schizotypy using partial least squares correlation. *Cortex*. 2016;81:137–150.

49. Modinos G, Mechelli A, Pettersson-Yeo W, Allen P, McGuire P, Aleman A. Pattern classification of brain activation during emotional processing in subclinical depression: psychosis proneness as potential confounding factor. *PeerJ*. 2013;1:e42.

50. Jeong JW, Wendimagegn TW, Chang E, et al. Classifying schizotypy using an audiovisual emotion perception test and scalp electroencephalography. *Front Hum Neurosci*. 2017;11:450.

51. Biswal B, Yetkin FZ, Haughton VM, Hyde JS. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn Reson Med*. 1995;34:537–541.

52. Smith SM, Miller KL, Salimi-Khorshidi G, et al. Network modelling methods for FMRI. *Neuroimage*. 2011;54:875–891.

53. Craddock RC, James GA, Holtzheimer PE III, Hu XP, Mayberg HS. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum Brain Mapp*. 2012;33:1914–1928.

54. Madsen KH, Churchill NW, Mørup M. Quantifying functional connectivity in multi-subject fMRI data using component models. *Hum Brain Mapp*. 2017;38:882–899.

55. Smith SM, Vidaurre D, Beckmann CF, et al. Functional connectomics from resting-state fMRI. *Trends Cogn Sci*. 2013;17:666–682.

56. Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp*. 2002;15:1–25.

57. Strother SC, Anderson J, Hansen LK, et al. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *Neuroimage*. 2002;15:747–771.

58. Guyon I. *Feature Extraction : Foundations and Applications*. Berlin, Heidelberg: Springer-Verlag; 2006.

59. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46:389–422.

60. Ho TK. Random decision forests. Third International Conference on Document Analysis and Recognition, {ICDAR} 1995, Montreal, Canada, 14–15 August 1995. Vol I. 1995:278–282.

61. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825–2830.

62. Abraham A, Pedregosa F, Eickenberg M, et al. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*. 2014;8:14.

63. Schrouff J, Rosa MJ, Rondina JM, et al. PRoNTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics*. 2013;11:319–337.

64. Hanke M, Halchenko YO, Sederberg PB, Hanson SJ, Haxby JV, Pollmann S. PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*. 2009;7:37–53.

65. Zou QH, Zhu CZ, Yang Y, et al. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. *J Neurosci Methods*. 2008;172:137–141.

66. Zang YF, He Y, Zhu CZ, et al. Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain Dev*. 2007;29:83–91.

67. Zang Y, Jiang T, Lu Y, He Y, Tian L. Regional homogeneity approach to fMRI data analysis. *Neuroimage*. 2004;22:394–400.

68. Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp*. 1995;2:189–210.

69. Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London, United Kingdom: Elsevier; 2007.

70. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage*. 2012;62:782–790.

71. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res*. 1996;29:162–173.

72. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*. 2002;15:273–289.

73. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*. 2006;31:968–980.

74. Goutte C, Toft P, Rostrup E, Nielsen F, Hansen LK. On clustering fMRI time series. *Neuroimage*. 1999;9:298–310.

75. Churchill NW, Madsen K, Mørup M. The functional segregation and integration model: mixture model representations of consistent and variable group-level connectivity in fMRI. *Neural Comput*. 2016;28:2250–2290.

76. Newman ME, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2004;69:026113.

77. Nowicki K, Snijders TAB. Estimation and prediction for stochastic blockstructures. *J Am Stat Assoc*. 2001;96:1077–1087.

78. Mørup M, Madsen KH, Dogonowski AM, Siebner H, Hansen LK. Infinite relational modeling of functional connectivity in resting state fMRI. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, eds. *Advances in Neural Information Processing Systems*, Vol 23; 6–11 December 2010; Vancouver, Canada. Curran Associates, Inc., Red Hook, NY; 2010:1750–1758. http://papers.nips.cc/paper/4057-infinite-relational-modeling-of-functional-connectivity-in-resting-state-fmri.pdf

79. Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*. 2009;10:186–198.

80. Cole DM, Smith SM, Beckmann CF. Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. *Front Syst Neurosci*. 2010;4:8.

81. Calhoun VD, Adali T, Pearlson GD, Pekar JJ. A method for making group inferences from functional MRI data using independent component analysis. *Hum Brain Mapp*. 2001;14:140–151.

82. Beckmann CF, Smith SM. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans Med Imaging*. 2004;23:137–152.

83. Cortes C, Vapnik V. Support-vector networks. *Kluwe Acad Publ*. 1995;297:273–297.

84. Vapnik V, Lerner A. Pattern recognition using generalized portrait method. *Autom Remote Control*. 1963;24:774–780.

85. Chang CC, Lin CJ. Libsvm a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2:1–27.

86. Schmah T, Yourganov G, Zemel RS, Hinton GE, Small SL, Strother SC. Comparing classification methods for longitudinal fMRI studies. *Neural Comput*. 2010;22:2729–2762.

87. Bishop CM. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ: Springer-Verlag New York, Inc.; 2006.

88. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16:412–424.

89. Rasmussen PM, Madsen KH, Lund TE, Hansen LK. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *Neuroimage*. 2011;55:1120–1131.

90. Haufe S, Meinecke F, Görgen K, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*. 2014;87:96–110.

91. Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik V. Support vector regression machines. In: Mozer MC, Jordan MI, Petsche T, eds. Advances in Neural Information Processing Systems 9, NIPS; December 2–5, 1996; Denver, CO. Cambridge, MA: MIT Press; 1997:155–161.

92. Plis SM, Hjelm DR, Salakhutdinov R, et al. Deep learning for neuroimaging: a validation study. *Front Neurosci*. 2014;8:229.

93. Jang H, Plis SM, Calhoun VD, Lee JH. Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: evaluation using sensorimotor tasks. *Neuroimage*. 2017;145:314–328.

94. Kim J, Calhoun VD, Shim E, Lee JH. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *Neuroimage*. 2016;124:127–146.

95. Guo X, Dominick KC, Minai AA, Li H, Erickson CA, Lu LJ. Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Front Neurosci*. 2017;11:460.

96. Vieira S, Pinaya WH, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci Biobehav Rev*. 2017;74:58–75.

97. Wang B, Klabjan D. Regularization for unsupervised deep neural nets. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17); February 4–9, 2016; San Francisco, CA. Ithaca, NY: Cornell University Library; 2017:1–6. http://arxiv.org/abs/1608.04426

98. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 27 (NIPS '14)*, Vol 27; December 8–13, 2014; Montréal, Canada. Curran Associates, Inc.; 2014:3320–3328. http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf

99. Xu Y, Jia R, Mou L, et al. Improved relation classification by deep recurrent neural networks with data augmentation. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics; December 13–16, 2016; Osaka, Japan. Ithaca, NY: Cornell University Library; 2016:1461–1470. http://arxiv.org/abs/1601.03651

100. Smith SM, Beckmann CF, Andersson J, et al.; WU-Minn HCP Consortium. Resting-state fMRI in the human connectome project. *Neuroimage*. 2013;80:144–168.

101. Biswal BB, Mennes M, Zuo XN, et al. Toward discovery science of human brain function. *Proc Natl Acad Sci USA*. 2010;107:4734–4739.

102. Churchill NW, Oder A, Abdi H, et al. Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. *Hum Brain Mapp*. 2012;33:609–627.

103. McIntosh AR, Bookstein FL, Haxby JV, Grady CL. Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage*. 1996;3:143–157.

C

Title

Classification of social anhedonia using temporal and spatial network features from a social cognition fMRI task

Authors

Krohne, Laerke G ;Wang, Yi; Hinrich, Jesper L; Moerup, Morten; Chan, Raymond CK; Madsen, Kristoffer H

Journal

Human Brain Mapping

Year

2019

RESEARCH ARTICLE

WILEY

Classification of social anhedonia using temporal and spatial network features from a social cognition fMRI task

Laerke Gebser Krohne[1,2,3,4] | Yi Wang[3,5] | Jesper L. Hinrich[1] |

Morten Moerup[1] | Raymond C. K. Chan[3,4,5] | Kristoffer H. Madsen[1,2,4]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

[2]Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital, Hvidovre, Denmark

[3]Neuropsychology and Applied Cognitive Neuroscience Laboratory, CAS Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

[4]Sino-Danish College, University of Chinese Academy of Sciences, Beijing, China

[5]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

Correspondence

Kristoffer H. Madsen, Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital, Kettegard Allé 30, 2650 Hvidovre, Denmark.
Email: kristofferm@drcmr.dk

Raymond C. K. Chan, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Beijing 100101, China.
Email: rckchan@psych.ac.cn

Abstract

Previous studies have suggested that the degree of social anhedonia reflects the vulnerability for developing schizophrenia. However, only few studies have investigated how functional network changes are related to social anhedonia. The aim of this fMRI study was to classify subjects according to their degree of social anhedonia using supervised machine learning. More specifically, we extracted both spatial and temporal network features during a social cognition task from 70 subjects, and used support vector machines for classification. Since impairment in social cognition is well established in schizophrenia-spectrum disorders, the subjects performed a comic strip task designed to specifically probe theory of mind (ToM) and empathy processing. Features representing both temporal (time series) and network dynamics were extracted using task activation maps, seed region analysis, independent component analysis (ICA), and a newly developed multi-subject archetypal analysis (MSAA), which here aimed to further bridge aspects of both seed region analysis and decomposition by incorporating a spotlight approach. We found significant classification of subjects with elevated levels of social anhedonia when using the times series extracted using MSAA, indicating that temporal dynamics carry important information for classification of social anhedonia. Interestingly, we found that the same time series yielded the highest classification performance in a task classification of the ToM condition. Finally, the spatial network corresponding to that time series included both prefrontal and temporal-parietal regions as well as insula activity, which previously have been related schizotypy and the development of schizophrenia.

KEYWORDS

archetypical analysis, decomposition, functional connectivity, social anhedonia, support vector classification

1 | INTRODUCTION

In the perspective of schizophrenia as a neurodevelopmental disease, it is very important to study potential early risk groups (Insel, 2010; Lewis & Levitt, 2002; Weinberger, 1987). Schizotypy refers to a set of positive, negative, or disorganized personality traits that are related to

schizophrenia (Ettinger et al., 2015). Individuals with schizotypy are nonclinical subjects, but they have some psychotic-like experiences, ranging from few (low schizotypy) to numerous (high schizotypy), which reflect their vulnerability for developing schizophrenia-spectrum disorders (Blanchard, Collins, Aghevli, Leung, & Cohen, 2011; Kwapil, 1998; Mason, 2015). The importance of studying schizotypy is twofold. Firstly, it has been suggested that early detection and intervention of schizophrenia might yield substantial improvements in treatment outcome, comparable to what has been reported in preventive approaches to cardiac death (Insel, 2010). Secondly, schizotypy studies have shown to increase the understanding of the psychopathology of schizophrenia.

Anhedonia, which is the reduced capability to experience pleasure in normal pleasurable situations, is considered as a negative dimension of schizotypy. High levels of anhedonia have consistently been reported in patients with schizophrenia (Blanchard et al., 2011; Bora, Yucel, & Pantelis, 2009) and ultra-high risk groups (Bora & Pantelis, 2013). Furthermore, longitudinal studies have shown that subjects with a high level of *social* anhedonia (reduced pleasure experience in social contexts) are more likely to develop schizophrenia-spectrum disorders later on, compared to control groups or high scorers on positive schizotypy (measured by perceptual aberration scale and magical ideation scale; Blanchard et al., 2011; Kwapil, 1998) (Gooding, Tallent, & Matts, 2005; Wang et al., 2014). For these reasons, social anhedonia will be the focus in this study.

On the other hand, the importance of social cognition research in understanding psychopathology of schizophrenia has been acknowledged (Green, Horan, & Lee, 2015; Penn, Sanna, & Roberts, 2007). Studies have shown that social cognition is substantially impaired in patients with schizophrenia and early risk groups (Bora & Pantelis, 2013; Fett, Viechtbauer, Dominguez M de, & Krabbendam, 2011), and changes have even been reported in subjects with schizotypy (Blanchard et al., 2011; Morrison, Brown, & Cohen, 2013). Theory of mind (ToM) is often defined as the ability to attribute mental states to ourselves and others, and consists of both a cognitive (centered about processing of knowledge and believes) as well as an affective (emotional processing) component (Sebastian et al., 2012; Shamay-Tsoory, Harari, Aharon-Peretz, & Levkovitz, 2010). The affective aspect is very similar to what is often defined as cognitive empathy (Sebastian et al., 2012), and will for simplicity, be referred to as empathy (Emp) in the rest of the article. The abnormalities of ToM or empathy ability has been related to schizotypy (Bora & Pantelis, 2013; Pickup, 2006). In particular, previous studies consistently suggested an association between high negative schizotypy and poor metalizing ability measured by self-report scales (Bedwell et al., 2014; Henry, Bailey, & Rendell, 2008; Thakkar & Park, 2010; Wang et al., 2013) and behavioral tasks (Pflum & Gooding, 2018; Thakkar & Park, 2010).

Functional imaging studies have correlated the degree of schizotypy and activity in isolated brain regions reviewed in (Ettinger et al., 2015; Nelson, Seal, Pantelis, & Phillips, 2013), however, until now, only relatively few studies have investigated how functional connectivity changes in individuals with schizotypy. Lagioia et al. determined six resting state networks and found that functional connectivity in the visual and auditory networks were correlated to the degree of schizotypy (Lagioia, Van De Ville, Debbané, Lazeyras, & Eliez, 2010). In terms of social anhedonia, studies found altered connectivity between the striatal seeds and the cingulate cortex as well as the insula during resting state (Wang et al., 2016) and altered functional connectivity of the amygdala during facial emotion processing task (Wang et al., 2018). Although previous studies have looked at correlations between brain activation or connectivity and the degree of schizotypy, actual classification is of great importance to determine if these changes can be used to categorize or even diagnose subjects already in early stages. Machine learning methods have been used in classification of schizophrenia patients from healthy control using functional imaging data (reviewed in (Madsen, Krohne, Cai, Wang, & Chan, 2018)). So far there are a few studies that have investigated the classification performance of individuals with schizotypy based on brain activation during task-based fMRI using machine learning methods (Modinos et al., 2012; Shinkareva, Ombao, Sutton, Mohanty, & Miller, 2006), but both studies only focused on the positive dimension of schizotypy instead of negative schizotypy.

The aim of our study was to investigate which features extracted from functional networks during a social cognition task were sufficient to classify subjects according to their degree of social anhedonia using supervised machine learning. To this end, we extracted brain network features using both standard activation maps and traditional seed region analysis (Biswal, Yetkin, Haughton, & Hyde, 1995; Cole, Smith, & Beckmann, 2010), but also decomposition methods based on independent component analysis (ICA; Beckmann & Smith, 2004; Calhoun, Adali, Pearlson, & Pekar, 2001) and the multi-subject archetypal analysis (MSAA) described in (Hinrich et al., 2016). Seed based analysis procedures extract features from defined seed regions, whereas ICA uses unsupervised learning to decompose the data into latent maximally independent spatial components. Each of these components can be thought of as representing a functional brain network. MSAA can be seen as seed region-based analysis where the seeds are automatically defined based on unsupervised learning. The features used for the classification, and the relation between the approaches are illustrated in Figure 1.
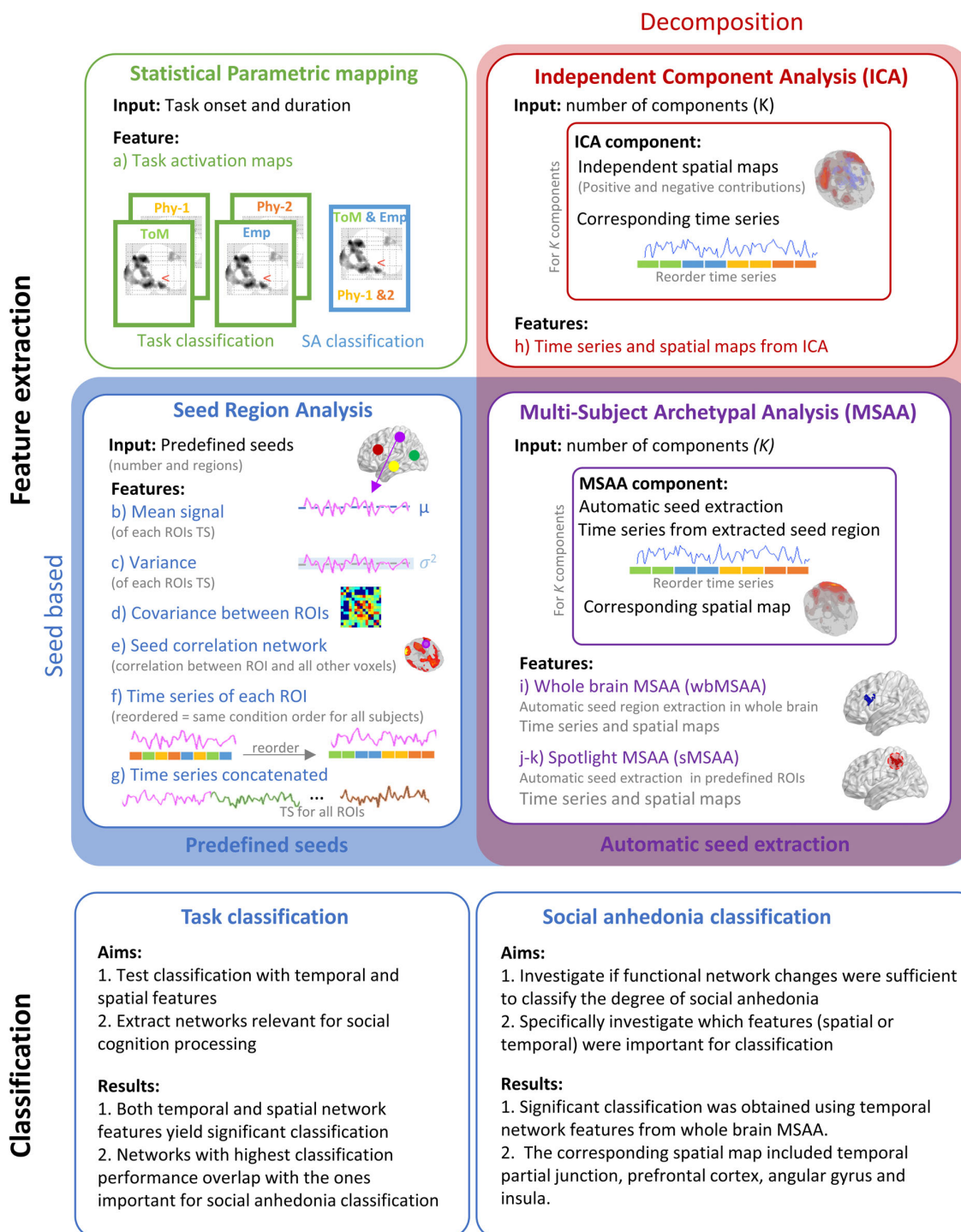
The use of different methods helped us exploring the separate importance of spatial and temporal network features.

Our second aim was to specifically investigate which features were important for classification. We investigated time series extracted from either specific brain regions or from networks, and hypothesized that the features showing significant classification of subjects with high social anhedonia would entail brain regions previously associated with schizotypy and the development of schizophrenia. Such regions include as prefrontal cortex, temporal–parietal regions, and insula (Chan, Di, McAlonan, & Gong, 2011; Kühn, Schubert, & Gallinat, 2012; Takahashi, Wood, Yung, Velakoulis, & Pantelis, 2009).

2 | MATERIALS AND METHODS

2.1 | Participants

This study included 76 college students from Guangzhou Medical University (37/39 male/female) with age between 17 and 21 years

FIGURE 1 Illustration of the feature extraction methods and aims of classification. We roughly divide the feature extraction methods considered into statistical parametric mapping, unsupervised decomposition, and seed region analysis. Here the letters a–k refers to the results of individual analyses as displayed in Table 1. (a) Refers to spatial maps extracted from statistical parametric mapping and classification approach (b,c) are based on static measures from seed based analysis, (d,e) are expressions of functional connectivity within and between the seeds and (f,g) reflect temporal dynamics of seed based analysis. In analyses (f–k) the time series are rearranged such that the order of the conditions is consistent across subjects, this was necessary as the order of the tasks were randomized across participants. In approach (h) time series and spatial maps obtain from ICA are considered, and approaches (i–k) are based on archetypical analysis which can be seen as seed based analysis with automatic extraction of seeds, merging aspects of ICA and seed region analysis. For approaches (e,f and h–k) classification was performed for each ROI/component separately, and thus multiple comparisons correction was used to assess the significance of the results [Color figure can be viewed at wileyonlinelibrary.com]

(μ = 19.3 years, σ = 0.9 years). The subjects were chosen such that they covered a continuous range of schizotypy and none had a history of drug abuse, or psychiatric disorders. The Chapman social anhedonia scale (CSAS) was used to assess the inability to experience pleasure from social interactions (Chan et al., 2015; Chapman, Chapman, Kwapil, Eckblad, & Zinser, 1994). The CSAS consists of 40 items (e.g., "Just being with friends can make me feel really good"; "Making new friends isn't worth the energy it takes") and higher score indicated more severity of anhedonia. The good reliability and validity of the CSAS has been proved in Chinese context (Chan et al., 2015). The internal consistency coefficient was 0.84 in our sample. The mean and standard deviation of all four Chapman scales and the Becks Depression Inventory can be found in Data S1–Table S2. All subjects were right-handed and a radiologist screened all scans to exclude subjects with any incidental clinical abnormalities. The study was approved by the Ethics Committee of the Institute of Psychology at the Chinese Academy of Sciences.

In a previous analysis, the same dataset showed specific correlation between the degree of social anhedonia and the mean activity in; the middle temporal gyrus, the temporoparietal junction and the medial prefrontal gyrus. (Wang et al., 2015). In contrast, this study investigated if the measured changes were sufficient for actual classification of subject with high and low social anhedonia (HSA/LSA) using support vector machines.

Subjects were defined in the HSA group if their CSAS score was more than one standard deviation above the mean (based on a large independent dataset including 887 subjects (Chan et al., 2012)). This separation threshold was relatively low, but comparable with what previously has been used in the literature (Wang et al., 2016). Furthermore, even when using this relatively low separation boundary, the dataset was unbalanced (HSA = 14/LSA = 56 subjects). As it will be discussed more carefully in Sections 2.9 and 3.4 this complicated the classification procedure.

2.2 | Functional imaging task

A Chinese adaption of the visual comic strip task developed by Völlm et al. was presented in a block design (Völlm et al., 2006; Wang et al., 2015). The task included four different conditions namely ToM, empathy, and two corresponding control conditions; "physical causality with one character" (Phy1) and "physical causality with two characters" (Phy2). Whereas the ToM and empathy condition were designed to probe the corresponding social cognition processing, the physical conditions were designed to look as similar to the social cognition conditions as possible. Hence, Phy1 included comic strips with only one character, whereas Phy2 included two interacting characters. Each condition was presented twice, resulting in a total of eight blocks, with each block containing five trials of comic strips belonging to the same condition. When the condition was presented the second time, a new set of comic strips were used, hence each comic strip was only seen once by each subject. In each trial, three pictures depicting a short story were displayed in the upper half of the screen for 6 s. Afterward, two pictures appeared in the lower half of the screen for

another 6 s. During the second 6-s period, participants were asked to choose one of the two pictures from the lower half of the screen as the appropriate ending to the story by pressing the corresponding button with their right hand. For the ToM trials, the original cartoons from the "Attribution of intention" (Brunet, Sarfati, Hardy-Baylé, & Decety, 2000) condition was used and the question: "What will the main character do next?" was asked. For the empathy condition, scenarios with emotional states attribution was showed and the question "What will make the main character feel better?" was asked. The total duration of the whole task was 8 min and 48 s. To control for effects of practice and fatigue the blocks were randomized across subjects. More details, as well as examples on the comic strip task, can be found in (Völlm et al., 2006), who developed the task.

2.3 | Image acquisition and preprocessing

All scans were acquired on a 3T Siemens Verio MR scanner at Guangzhou First People's Hospital in 2012, using a T2* weighted gradient echo based echo planar imaging (EPI) sequence with echo time = 28 ms, repetition time = 2,000 ms and flip angle = 90°. 264 whole brain volumes were acquired with a slice thickness of 4 mm, matrix size 64×64 (32 slices in coronal plane), field of view = 210×210 mm, voxel size = $3.3 \times 3.3 \times 4$ mm, and bandwidth = 2,232 Hz/px.

The images were preprocessed using Statistical Parameter Mapping (SPM) version 12 revision 6685. The eight first volumes of the scans were removed to ensure T1 equilibrium, and slice-timing correction was performed to correct for the descending slice order with the middle slice as reference. The EPI images were normalized to the EPI template (ICBM-152) and the images were re-sliced to $3 \times 3 \times 3$ mm. As this study focused on functional connectivity modeling additional preprocessing steps were included, since artifacts can lead to spurious connections (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012). First despiking was performed to remove transient phenomena without scrubbing (Patel et al., 2014) using a Daubechies 4 mother wavelet. Then additional nuisance regressors were included in a multiple linear regression and the effect of them was removed from the data. These included; (a) mean signal and second order detrending (b) nuisance variable regressors (NVRs), (c) spike percentage from despiking, and (d) explicit modeling of specific time frames based on the DVARS and frame wise displacement criteria as described in (Patel et al., 2014; Power et al., 2012), using a threshold of 1% and 1 mm, respectively. NVRs were used to remove both *residual motion* (24-parameter Volterra expansion model (Friston, Williams, Howard, & Frackowiak, 1996) based on the six head motion parameters estimated during realignment) and *physiological noise* where the mean signals from nonneuronal brain regions was extracted. Nonneuronal tissue included white matter, which was segmented using the SPM12 tissue probability map with a threshold of 0.5, cerebrospinal fluid in the lateral ventricles according to the HarvardOxford atlas (Desikan et al., 2006). To reduce the influence of partial volume effect with gray matter, the white matter mask was eroded by two voxels. Finally, the images were smoothed using an isotropic Gaussian 8 mm full width at half maximum filter.

2.4 | Classification using support vector classification

To classify subject into high and low social anhedonia, as well as for the task classification, we used binary support vector machines to perform supervised classification (Cortes & Vapnik, 1995). The goal of support vector classification (SVC) is to identify a function that discriminates the labels (e.g., high or low social anhedonia) in a training dataset, such that it is possible to use this function to classify the labels of a test dataset. In principle, it is possible to apply SVC directly to the (preprocessed) fMRI images. However, due to the very high dimensionality of fMRI images in relation to the number of subjects, perfect classification in the training dataset is trivial but with poor generalization to the test data due to overfitting (see Madsen et al. for a more thorough description of SVC for fMRI data (Madsen et al., 2018)). We therefore applied SVC on 11 spatial and temporal features (analysis a-k listed in Table 1), which were extracted from the fMRI data to capture the network changes of interest.

In short, one feature included the task specific activation maps determined by a SPM analysis (Section 2.5), six features resulted from a seed region analysis (Section 2.6), and four came from the decomposition methods (Section 2.7). For some of the seed region analysis and decomposition methods, we extracted both time series and spatial maps for each seed region/component respectively (analysis e,f and h–k), and classification was then performed on each extracted feature respectively. Table 1 lists the classification performances of the features yielding the highest classification performance, and maximum permutation statistics was therefore used to correct for multiple comparisons between the components as described in Section 2.9.

For classification, we used the SVC-C implementation from the LIBSVM (Chang & Lin, 2011) library with a linear kernel. We used nested cross validation to determine the soft margin penalty parameter, and to evaluate the classification performance. For task classification, the cross validation scheme was based on grouped stratified cross validation where each subject was considered a group. In the inner loop, the optimal soft margin penalty parameter (C-parameter) was determined in a logarithmic grid containing 11 values $C \in [2^{-5}, 2^{-3}, ..., 2^{15}]$ by 10-fold cross validation, and an unbiased estimate of the classification accuracy was obtained in another outer 10-fold cross validation loop.

For HSA classification, a similar scheme was followed but without grouping as there was only one sample per subject in this case. Furthermore, the C-parameter was adjusted for each class to counteract the class imbalance (Chang & Lin, 2011). The inner and outer loops where set to reserve exactly one sample of the least common class (HSA) resulting in 13- and 14-fold cross validation, this ensured that stratification across splits was achievable while preserving sufficient data for training.

2.5 | Statistical parametric mapping

To determine task specific activity maps for all four task conditions (ToM, Emp, Phy1, and Phy2), we ran a standard SPM analysis, performing a parametric statistical test for each voxel separately. The significance level was $\alpha_{RFT} \leq 0.05$, where random field theory was used to correct for multiple comparisons. The activation maps were later used as features (classification approach [a]) for classification. Since the activation maps were constructed based on information about task onset and duration, we expected that they would obtain a high performance for classifying the tasks conditions. However, for the social anhedonia classification, which was not directly related to the presented task, the static nature of this feature extraction step might not identity information useful for classification. For the task classification, we used one task activation map for each social construct, that is, the ToM—Phy1 condition, and empathy—Phy2 condition, respectively. For the HSA classification, we used one single contrast map, reflecting the pooled effect of ToM, and empathy in comparison to the physical control conditions, as illustrated in Figure 1.

Furthermore, we used SPM to perform a pooled condition analysis (PCon) identifying the pooled effect of the social cognition tasks (ToM and Emp) compared to the control conditions. This was used as input for the spotlight MSAA as described in Section 2.7.

2.6 | Seed region analysis

Seed region analysis is a very intuitive way to investigate the brain by determining the activity in predefined regions of interest (ROIs). In this study, six different methods (approach b–g) were used to investigate the ROI specific activity, which later were used for classification. These included; approach (b): the mean activity and (c) variance within each ROI, (d) the covariance between all N ROIs (calculated pairwise), (e) the correlation between the time series of each ROI with all voxels in the brain (classical seed based analysis) resulting in a connectivity map for each seed, (f) the extracted time series of each ROI separately, and (g) the time series of all ROIs concatenated. All of these are illustrated in Figure 1, and enabled us to study the importance of temporal dynamics (approach (f) and (g)), network coupling (approach (d) and (e)), and static features separately.

The time series of each ROI were extracted as the first eigenvariate, which reflects the most consistent source across all included voxels. Compared to using the average across the ROI, this can be an advantage if there are multiple sources in the given ROI (Poldrack & Gorgolewski, 2014). When using the time series as feature for classification, they were rearranged (by simple temporal reordering) such that they reflect the same structure (ToM, Emp, Phy1, and Phy2) for all subjects, despite that the order of the conditions were randomized across subjects. In approach (e) the correlation between the time series of the ROIs, and that of all other voxels in the brain, was determined using Pearson's correlation coefficient, followed by conversion to Z-score through the Fisher Z-transform (Fisher & Fisher, 1915).

In approach (e) and (f) classification was performed independently for each ROI, highlighting the importance of multiple comparisons correction as described more carefully in Section 2.9.

TABLE 1 Classification performance of both task and HSA

		Description of classification feature	Task classification accuracy (%; p value)		HSA vs. LSA classification MCC (p-value)
			ToM--Phy1	Emp--Phy2	
Seed region analysis features					
	(a) Task activation maps	Task specific activation maps determined using SPM	84% $p = .001$ $1 \times V$	81% $p = .001$ $1 \times V$	0.13 $p = .199$ $1 \times V$
Static measures	(b) Mean activity	Average activity of each ROI	41% $p = .801$ $1 \times K$	56% $p = .115$ $1 \times K$	-- (†)
	(c) Variance	Variance within each ROI	58% $p = .070$ $1 \times K$	58% $p = .091$ $1 \times K$	−0.02 $p = .569$ $1 \times K$
Network coupling	(d) Covariance (network coupling)	Covariance of the time series of ROIs	60% $p = .039$ $1 \times (K^2+K)/2$	60% $p = .037$ $1 \times (K^2+K)/2$	0.43 $p = .005$ $1 \times (K^2+K)/2$
	(e) Seed based network	Correlation between time series of a ROI and all voxels in the brain	73% $p = .001$ $K \times V$	73% $p = .001$ $K \times V$	0.19 $p = .897$ $p_{UC} = .125$ $K \times V$
Time series	(f) Time series (ROI specific)	Time series of each ROI separately	59% $p = .666$ $K \times T_1$	61% $p = .393$ $K \times T_1$	0.35 $p = .189$ $p_{UC} = .007$ $K \times T_2$
	(g) Time series (concatenated)	Time series of each ROI, concatenated	63% $p = .010$ $1 \times KT_1$	68% $p = .001$ $1 \times KT_1$	−0.15 $p = .937$ $1 \times KT_2$

Decomposition features

		Feature type			
		TS		TS	SM
(h) ICA	Time series and spatial maps from ICA	73% $p = .001$ $K \times T_1$	79% $p = .001$ $K \times T_1$	0.45 $p = .072$ $p_{UC} = .005$ $K \times T_2$	0.24 $p = .912$ $p_{UC} = .093$ $K \times V$
(i) wbMSAA	Time series and spatial maps from wbMSAA	74% $p = .001$ $K \times T_1$	69% $p = .002$ $K \times T_1$	0.56 $p = .008$ $p_{UC} = .002$ $K \times T_2$	0.42 $p = .097$ $p_{UC} = .006$ $K \times V$
(j) sMSAA_Lit	Time series and spatial maps from spotlight MSAA (using literature coordinates)	67% $p = .020$ $K \times T_1$	73% $p = .001$ $K \times T_1$	0.49 $p = .032$ $p_{UC} = .003$ $K \times T_2$	0.25 $p = .744$ $p_{UC} = .059$ $K \times V$
(k) sMSAA_PCon	Time series and spatial maps from spotlight MSAA (using PCon coordinates)	-- (*)	--(*)	0.31 $p = .463$ $p_{UC} = .030$ $K \times T_2$	0.25 $p = .732$ $p_{UC} = .066$ $K \times V$

Note: For each performed analysis, this table yields a short explanation of the input feature and classification performance measured in accuracy (task classification) or Mathews correlation coefficient, MCC (HSA classification). For the HSA classification, both time series (TS) and spatial maps (SM) were used as features for the decomposition methods. For seed region analysis features e–f and decomposition methods (h–j) the table lists the classification performance of the component yielding the highest classification performance. The p-value was nonparametrically estimated with random permutation testing and maximum permutation statistics was used to correct for multiple comparisons when necessary. The number of comparisons × feature dimensionality are stated for each of the classification models, where the size of the voxel dimension is $V = 60,704$, $T_1 = 60$ (time points for each condition), $T_2 = 264$ (total number of time points), and $K = 25$ (number of components or ROIs). The uncorrected p-value (p_{UC}) was also based on random permutation and is stated for some HSA classifications. (†) HSA was not classified as the overall mean per subject was subtracted during preprocessing. (*) task classification was not calculated for the sMSAA_PCon analysis, since this result would be biased.

Abbreviations: Emp, empathy; HSA, moderately high social anhedonia; LSA, low social anhedonia; MCC, Matthews correlation coefficient; ToM, theory of mind.

2.7 | Decomposition methods

One of the most frequently used decomposition methods in neuroscience is the *ICA*, which determines a predefined numbers of maximally independent sources (McKeown et al., 1998). For fMRI data, these sources represent spatial networks, where all included regions have similar time series. For multi-subject analysis, common spatial components can be obtained by concatenating subject data in time (Calhoun, Adali, & Hansen, 2003). More specifically, ICA seeks to identify latent sources in the data from multiple mixed measurements via the per subject linear mixing model

$$X_i = A_i S_i + E_i,$$

where $X_i \in \mathbb{R}^{T \times V}$ is the data matrix measured at T timepoints and across V voxels for the i'th subject, $A_i \in \mathbb{R}^{T \times K}$ contains K source time series as columns, $S_i \in \mathbb{R}^{K \times V}$ is comprised by the K spatial components as rows, and $E_i \in \mathbb{R}^{T \times V}$ is a residual error term. While the expression above enforces no coupling across subjects, such dependence is usually accomplished by enforcing dependence or equality of S_i across subjects, which we will consider later. Since minimizing the residual leads to rotational ambiguity and thereby nonunique solutions, additional assumptions, or constraints are typically imposed on either the time series or spatial components or both. In spatial ICA, this typically amounts to assuming a non-Gaussian source distribution upon the spatial components.

MSAA is another data driven approach, which bridges aspects of seed analysis and decomposition (Hinrich et al., 2016) (Cutler & Breiman, 1994; Mørup & Hansen, 2012). MSAA is a latent variable model, similar to ICA, but is constrained to have latent factors that reflect representative points in the data, termed "archetypes." For fMRI data, the archetypes are a set of representative time-series, which have a corresponding set of spatial networks. Whereas ICA represents the fMRI data by a linear mixture of maximally independent spatial maps, MSAA determines the components through iterative optimization of; (a) a seed region matrix, C (that is identical for all subjects) and (b) a set of subject specific spatial maps (S) corresponding to each archetype. The archetypes for each subject are given as the weighted average of the voxels specified in the seed region matrix, such that

$$A_i = X_i C$$

where X_i is the subject specific data and $A_i \in \mathbb{R}^{T \times K}$ includes all archetypes defining distinct temporal profiles for the i'th subject. Figure 2 illustrates how MSAA represents the fMRI data as archetypes and spatial maps. Each voxel time series is reconstructed by convex combinations as defined in S_i of the archetypes. Thus, both the columns of S_i and C are constrained to be nonnegative and to sum to one. The resulting spatial maps can therefore be interpreted as the fractional contribution of all voxels to the archetypal time series as specified in A_i.

The MSAA decomposition is in general unique (Mørup & Hansen, 2012) and the linear model (per subject) can be formulated as

$$X_i = X_i C S_i + E_i,$$

Under the assumption of independently distributed additive Gaussian noise with heteroscedasticity over voxels we have

$$e_{i,v} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{i,v}^2\right),$$

Where $e_{i,v}$ is a time vector of the residual in voxel v for subject i and $\sigma_{i,v}^2$ is the voxel and subject specific noise variance. This lead to the likelihood
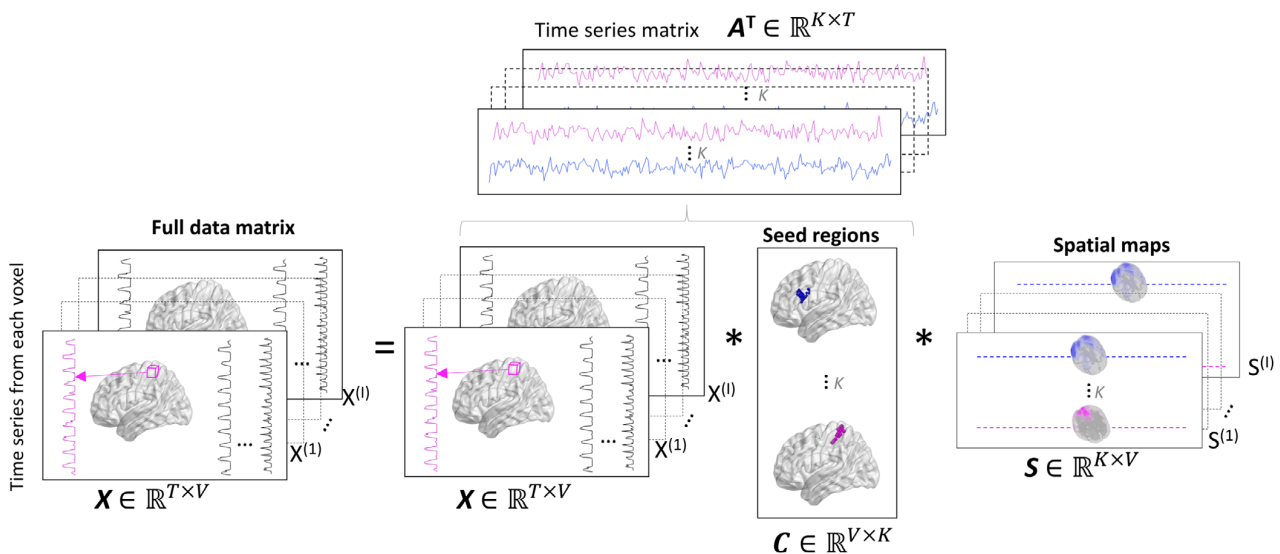
FIGURE 2 Illustration of whole brain multi-subject archetypal analysis (wbMSAA). The columns data matrix X include the time series for all V voxels. Through iterative optimization, the MSAA algorithm determines a seed region matrix C, specifying the optimal choice of K seed regions across subjects, as well as a set of K temporal (X_i C) and spatial components S_i for all B subjects. The model also includes a subject specific noise map, which is not specified in this figure [Color figure can be viewed at wileyonlinelibrary.com]

$$\mathcal{L} = \prod_i^B \prod_v^V \frac{1}{\left(2\pi\sigma_{i,v}^2\right)^{T/2}} \exp\left(-\frac{\|\mathbf{x}_{i,v} - \mathbf{X}_i \mathbf{C}\mathbf{s}_{i,v}\|^2}{2\sigma_{i,v}^2}\right).$$

Optimizing this likelihood leads to a sparse seed region matrix **C**, which selects the archetypical voxel time series that best span the entire dataset, and a corresponding set of subject specific spatial maps S_i. For explicit derivation of update rules see (Hinrich et al., 2016). Determining **C**, S_i, and σ_i is a nonconvex optimization problem (Mørup & Hansen, 2012), but a solution can be found by alternating optimization, that is, optimizing for **C** while keeping S_i and σ_i fixed and vice versa.

2.7.1 | Connection between ICA, seed based analysis, and MSAA

In the following, we show how the decomposition scheme of MSAA can be used to bridge spatial group ICA with seed based analysis. The MSAA directly finds subject specific spatial maps (S_i) and temporal activations (X_iC) which through the common seed matrix (C^{MSAA}) express variability across subjects. In contrast, spatial group ICA assumes the spatial sources are fixed across subjects (Calhoun et al., 2003), however, individual subject expressions (spatial maps) can be identified through either back reconstruction or dual regression (Erhardt et al., 2011). When the spatial sources are known and no additional constraints are imposed upon the time series, solving for A_i reduces to an ordinary least squares regression problem where the solution can be expressed as

$$A_i = X_i \bar{S}^\top \left(\bar{S}\bar{S}^\top\right)^{-1}.$$

Here $\bar{S}$ represents the shared spatial components. In back reconstruction, individual subject components are formed through the expression

$$X_i = A_i \widetilde{S}_i,$$

where $\widetilde{S}_i$ is the individual spatial components and inserting the expression for A_i we obtain

$$X_i = X_i \bar{S}^\top \left(\bar{S}\bar{S}^\top\right)^{-1} \widetilde{S}_i,$$

which again allows the individual spatial maps to be formed by solving an ordinary least squares problem. This establishes an attractive correspondence between MSAA and group ICA, where, in this case, the nonsparse "seed matrix" given by $C^{ICA} = \bar{S}^\top \left(\bar{S}\bar{S}^\top\right)^{-1}$ can take on both positive and negative values whereas the columns are not constrained to sum to one.

2.7.2 | Spotlight MSAA

In this study, we considered an expansion to the MSAA algorithm by implementing a spotlight approach that restricted the seed region

matrix to prespecified ROIs. This allowed specifying a subset of voxels from which the seed regions were then defined,

$$X_i = \widetilde{X}_i C S_i + E_i,$$

where $\widetilde{X}_i$ is the subset of voxel time series in the ROIs as illustrated in Figure 3. This approach is useful to investigate "archetypal generating activity" in specific areas, or if only approximate ROIs are known. The derivation is given in (Hinrich et al., 2016), though they did not investigate the restricted method or considered the stability of its solution.

In the remaining manuscript, we will refer to the restricted MSAA as spotlight MSAA (sMSAA) in contrast to the original whole brain MSAA (wbMSAA).

We have run two sMSAA analysis using seed region restriction maps from; (a) a literature study (sMSAA_Lit) and (b) from a pooled condition analysis (sMSAA_PCon) respectively, as described in Section 2.8.

Implementation

We applied group ICA through the GroupICATv4.0a GIFT toolbox (Rachakonda, Egolf, Correa, Calhoun, & Neuropsychiatry, 2015), using the Infomax algorithm and the corresponding default settings. The number of components was selected using the minimum description length as proposed in Li, Adali, and Calhoun (2007), which for our dataset resulted in 25 components. Finally, subject specific spatial and temporal components were determined using the default back reconstruction method implemented in GIFT (Calhoun et al., 2003). For visualization purposes, the spatial components where z-scored and both positive and negative contributions were shown.

For the MSAA analysis, we used the same number of components as for ICA. As the MSAA algorithm is a nonconvex optimization problem, there was a risk that the solution would get stuck in a local and not global minimum. As done for other nonconvex problems, we therefore repeated the analysis several times with different random initializations for each run, and chose the solution with the lowest final cost at the end of the optimization. Optimization halted after either a maximum of 250 iterations or when the relative decrease in the cost function was less than 10^{-6} as in Hinrich et al. (2016). Different initializations, such as the FurthestSum initialization (Mørup & Hansen, 2012) have been suggested for archetypal analysis. However, as these resulted in a higher final cost function, random initialization was used in this study.

To increase the stability of MSAA the algorithm was rerun with 10 random initializations choosing the solution that obtained the lowest cost function. To further investigate the stability of the algorithm we repeated the fitting procedure 10 times and compared the spatial maps across runs using spatial correlation, this indicated that components were fairly stable across runs, providing an average correlation of 0.86. Visual inspection revealed that the differences were primarily due to minor changes in network expressions between runs for some components, see the stability of wbMSAA section in Data S1 for further information. Furthermore, the finding of significant classification of HSA using wbMSAA times series reproduced in all 10 individual runs.
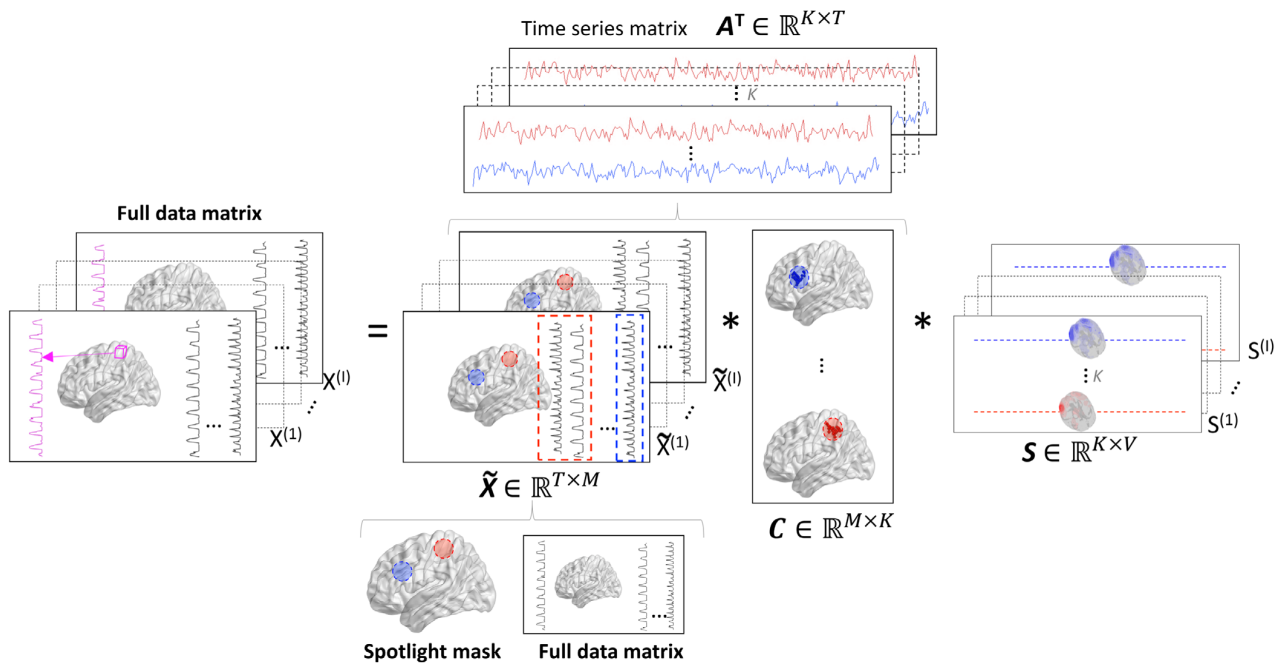
FIGURE 3 Illustration of the spotlight (sMSAA) approach. For the spotlight MSAA **C** and **X** are restricted to only include a subset of the voxels corresponding to some predefined regions of interest (for simplicity only two regions are shown here). However, the exact localization and size of the seed regions are still optimized by the algorithm. Apart from the restriction, the model is identical to the wbMSAA shown in Figure 2 [Color figure can be viewed at wileyonlinelibrary.com]

2.8 | Predefined ROIs

For the seed regions, analysis and spotlight MSAA predefined ROIs were a prerequisite for the analysis. We defined the ROIs as all voxels in a sphere (8 mm radius) around a given center coordinate. These were determined through a literature study of ToM and empathy processing, taking into account both reproducibility of the areas (Abu-Akel & Shamay-Tsoory, 2011; Shamay-Tsoory et al., 2010) and specificity for the comic strip task (Benedetti, Bernasconi, Bosia, & Smeraldi, 2009; Völlm et al., 2006; Wang et al., 2015). The center coordinates are illustrated and labeled in Figure 4 and the MNI coordinates can be found in Data S1–Table S3.

Finally, for the classification of social anhedonia using spotlight MSAA, center coordinates were also obtained using the peak coordinates of significant clusters for the pooled condition analysis (PCon) as described in Section 2.4. All center coordinates can be found in Data S1–Table S3.

2.9 | Statistical tests and measures

We used the accuracy as performance measure for the task classification, as it provides a straightforward interpretation for balanced samples. However, for the classification of unbalanced datasets the accuracy measure can be misleading. That is, even in the case of a trivial classification where all subjects were classified as the dominant class (e.g., in this study: LSA = 56, HSA = 14), the accuracy would be 56/(56 + 14) = 80%. To mitigate this issue, we used the Matthews

correlation coefficient (MCC) for the social anhedonia classification, as it is regarded as being one of the best summary statistic measures for unbalanced datasets (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000; Powers, 2011). MCC returns a value between −1 (worst) and 1 (best) where 0 indicates that the result is no better than random classification.

For all classification procedures, statistical inference of the performance was performed using a random permutation testing procedure (Nichols & Holmes, 2003). For each of 1,000 random permutations the entire classification procedure, including the inner and outer nested cross validation loops, were repeated to obtain an empirical null distribution of the performance measure (accuracy and MCC for task and HSA classification respectively).

As mentioned above, for some features the classification was performed for each ROI/network separately, and the significance of these analyses therefore needed to be corrected for multiple comparisons. This was done by the use of maximum permutation statistics, where an empirical null distribution was obtained by considering only the most significant effect over the entire set (here regions or components), which controls the family-wise error over the set.

3 | RESULTS AND DISCUSSION

This combined results and discussion section is split into five subsections, covering different aspects of the study. The first section includes a general discussion of the networks determined
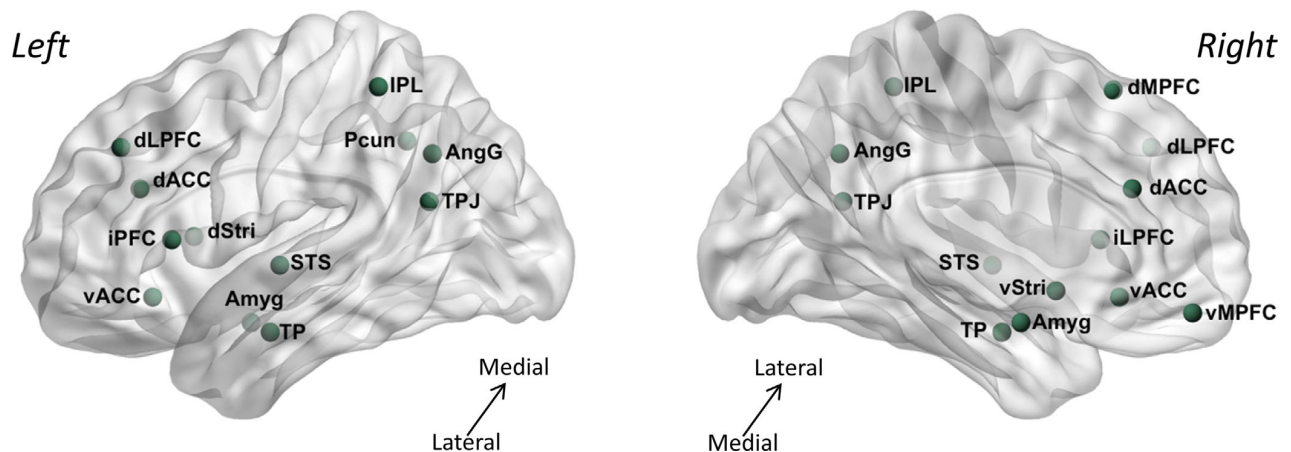
FIGURE 4 Illustration of center coordinates determined based on the literature. These nodes were used both for the seed region analysis approaches, and for the spotlight MSAA. Abbreviations: Amyg, amygdala; AngG, angular gyrus; d/v ACC, dorsal/ventral anterior cingulate cortex; d/v mPFC, dorsal/ventral medial prefrontal cortex; d/v Stri, dorsal/ventral striatum; IPL, inferior parietal lobule; i/dL PFC, inferior/dorsolateral prefrontal cortex; Pcun, precuneus; STS, superior temporal sulcus; TP, temporal pole; TPJ, temporoparietal junction [Color figure can be viewed at wileyonlinelibrary.com]

by the decomposition methods (ICA and MSAA), and comments on the stability of these approaches. Sections 3.2 and 3.3 cover the results from the task and social anhedonia classification, respectively, and discuss how these findings correspond to our hypotheses and previous literature. Since MSAA is a new decomposition method, which previously only has been applied in one neuroimaging study (Hinrich et al., 2016), we comment on the general interpretability and stability of the MSAA networks, and compare it with ICA in Section 3.4. Finally, in Section 3.5 we discuss general limitations of our study, as well as suggestions for future development and applications.

3.1 | Network extraction using decomposition methods

Visual inspection of the spatial maps from ICA and MSA showed that both methods captured networks which previously have been related to ToM processing (Benedetti et al., 2009; Völlm et al., 2006; Wang et al., 2015), without any a priori knowledge about the task onset and duration (which was a requirement for the previous studies that used SPM analysis). Furthermore, we observed that both ICA and MSAA successfully captured effect of no interest (such as pulsation and movement artifacts) as well as other specific activity (visual or motor processing) in separate networks. This is an important sanity check, as noise/unrelated activity would otherwise contaminate the task related networks.

3.1.1 | Stability

As described in Section 2.7, the wbMSAA algorithm was run 10×10 times, comparing the stability of the spatial networks, when the best (lowest final cost) solution of 10 runs was compared

for 10 repetitions. Using greedy matching a mean correlation of 86% was obtained. Visual inspection showed that the same networks were found in all 10 runs, but with minor differences, resulting in the nonperfect matching. Using the 10 repeated runs to investigate the classification stability, the same feature (discussed later in Section 3.3) was found to result in the highest classification performance (MCC varied between 0.49 and 0.56), which was significant for all 10 repetitions. This stability analysis was only performed for the wbMSAA. For the spotlight approaches the algorithm was repeated 10 times, and the solution with the lowest cost function was chosen.

3.1.2 | Cross validation

We used stratified k-fold cross validation as described in Section 2.4. For cross validation, it is important that the test and training data sets are independent. For the seed region analysis features, this is naturally the case, as the feature extraction was performed for each subject separately. However, in order to limit the computational complexity and to ensure correspondence of components across cross validation splits for ICA and MSAA, the decomposition was run on the entire dataset. Note that this did not lead to biased estimates of the classification performance, as no information about the class labels were used in the decomposition step.

3.2 | Classification of task conditions

The aim of the task classification was twofold. Firstly, it was a proof of concept of our classification approach, using either temporal or spatial network features as input to the SVC. Secondly, we wanted to investigate if the information captured by the networks was sufficient to actually classify task conditions, and to see how the networks

important for classification correspond to previous literature on ToM and empathy processing. The classification performances are listed in Table 1, and networks are illustrated in Figure 5.

First, we used the activation maps from the *SPM analysis* for classification. These activation maps yielded the highest task classification performance (mean accuracy of 83%), which was expected since they were informed about the onset and duration of the task conditions. The center coordinates, cluster size, and *z*-score of the significant clusters can be found in Data S1–Table S3. This result was mainly used to validate that there was sufficient signal difference between the task conditions.

To investigate our hypothesis about the importance of both temporal and spatial network dynamics, we used six features from the *seed region analysis* as illustrated in Figure 1. Firstly, we found that classification was not significant when using static measures such as the mean and variance, indicating that these simple measures do not capture enough signal difference between task blocks for classification in the considered sample. On the contrary, all *spatial networks features* (covariance and seed based analysis) resulted in significant classification with accuracies from 60 to 73%. As described in Section 2.7, classification was performed for each of the 25 networks extracted in the seed based analysis. Table 1 and Figure 5 include the
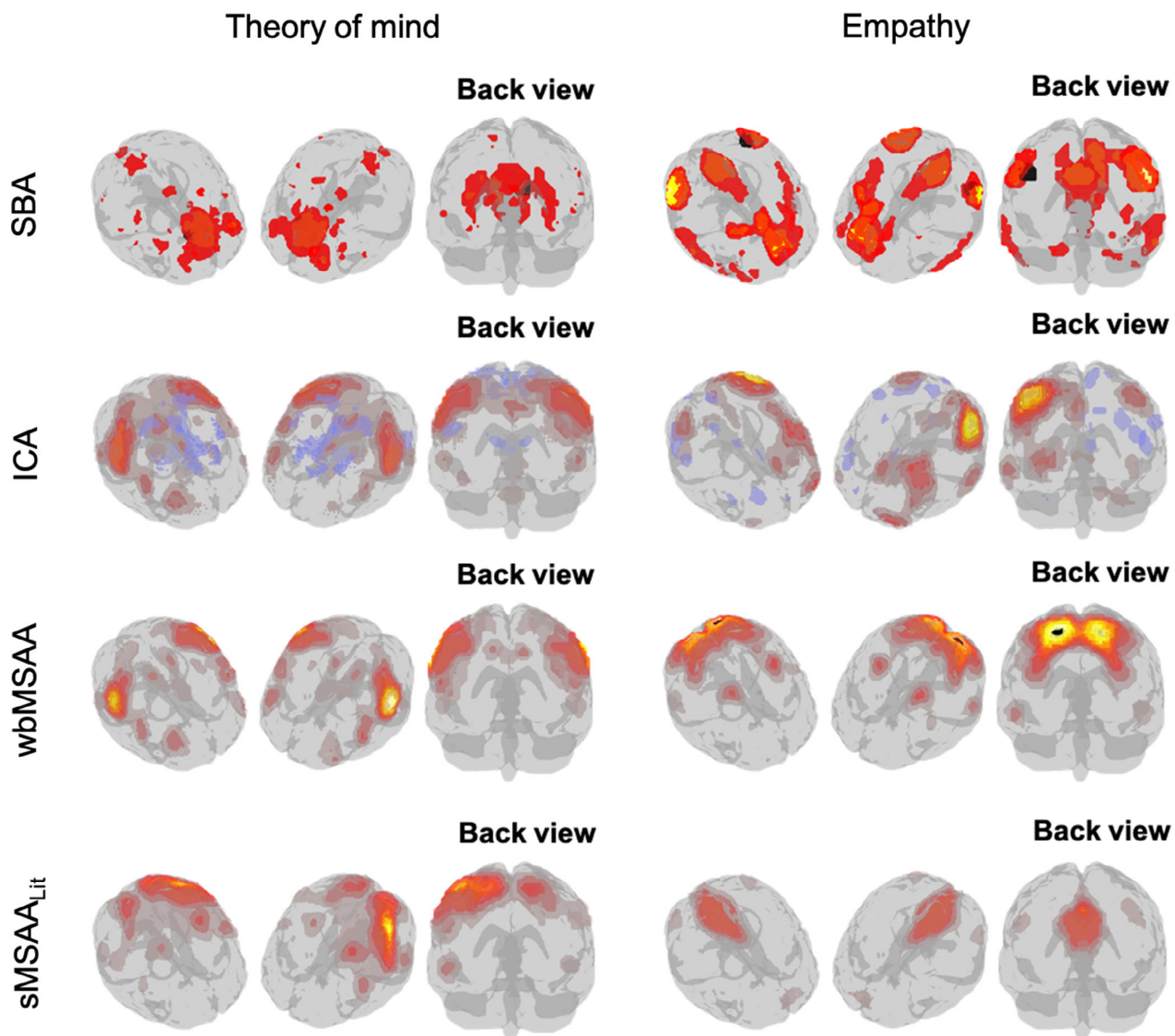
FIGURE 5 Mean spatial maps across subjects of the networks for ToM-Phy1 classification (left) and Emp-Phy2 classification (right), for SBA, ICA, wbMSAA, and sMSAA$_{lit}$, respectively. More significant networks can be found in Data S1–Figures S2–S4. For all four methods, the ToM-Phy1 classifying networks have most activity in the temporoparietal regions, and prefrontal regions. For the Emp-Phy2, processing similar regions are included, but generally more activity is located in posterior parietal regions. For visualization, the SBA networks include the most significant 10% of the network correlations, each ICA map was *z*-scored and thresholded at $Z = 1$, and the MSAA networks include voxels with 10% or more fractional contribution [Color figure can be viewed at wileyonlinelibrary.com]

networks that obtained the highest classification performance, however, more networks yielded significant classification, which can be found in Data S1–Figure S1. For the empathy condition (Emp-Phy2), the seed of the network yielding the highest classification performance was located in the angular gyrus and the network further included the inferior parietal lobule (IPL), precuneus, medial temporal gyrus, and medial prefrontal cortex (mPFC). For the ToM classification, the seed was located in the dorsal anterior cingulate cortex (ACC), and the network included frontal lobe regions, caudate and the precuneus. Most of these regions were suggested to be involved in the ToM or empathy related processing in previous studies (Abu-Akel & Shamay-Tsoory, 2011; Fan, Duncan, de Greck, & Northoff, 2011). In particular, the IPL, precuneus, middle temporal gyrus, mPFC, and ACC are key regions of default mode network, which plays important role in social processing, such as understanding others' beliefs and feelings and self-referencing (Andrews-Hanna, Reidler, Sepulcre, Poulin, & Buckner, 2010; Takeuchi et al., 2014).

Finally, we also tried to classify the conditions using the time series from the 25 seed regions. Here significant classification was only obtained when concatenating the time series from all components (mean accuracy 65%), and not when using the TS from each seed region separately.

For the *decomposition methods*, we used the time series from each component extracted using the three methods: ICA, wbMSAA, and sMSAA$_{Lit}$. All decomposition time series yielded a high classification performance with accuracies ranging from 67 to 79%, which were significant after correcting for multiple comparisons. The reason for the high classification performance when using time series from decomposition methods compared to seed region analysis, might be that the decomposition methods extract components which maximally explain the data. They therefore captured networks (and corresponding time series) which were the most prominent in the data, whereas, the seed region analysis relied on seed region points that were manually chosen based on previous literature, and thus were not specific for the given dataset.

The corresponding spatial maps of the best components are shown in Figure 5, and other significant networks can be found in Data S1–Figures S2–S4. Generally, we found that the best networks across most methods included similar regions. For the *ToM-Phy1* classification (left column), the networks include inferior and medial frontal gyrus, temporoparietal junction (TPJ), posterior cingulate cortex (PCC), and postcentral gyrus activation, which all are known to be involved in ToM processing (Amodio & Frith, 2006; Ettinger et al., 2015; Frith & Frith, 2006; Pickup, 2006). For the *Emp-Phy2* classification, the networks included similar regions as for the ToM-Phy1 classification, but generally there was more activation in posterior parietal regions, such as precuneus and PCC.

To summarize, our findings show that both spatial networks and temporal dynamics capture important information, which enabled significant classification of the ongoing social cognition task. The networks, which yielded the highest classification performance, generally included temporoparietal and prefrontal areas, which consistently have been

considered core regions for ToM and empathy processing (Frith & Frith, 2006; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014).

3.3 | Classification of social anhedonia

In this section, we show and discuss the results from the social anhedonia classification. The classification performances, measured by the MCC are listed in Table 1 and Figure 6 shows the spatial maps of the features obtaining the highest classification performance.

Whereas the activation maps from the SPM analysis resulted in the highest task classification performance of all methods, our results show that neither the raw maps, nor the seed based static measures (mean and variance) enabled significant classification of social anhedonia. In fact, for the *seed region analysis* features, only the covariance feature obtained significant classification with a MCC = 0.43 (p = .005). This indicates that simple network coupling between regions that are known to be involved in social cognition processing, seems to capture important information to differentiate the high and low social anhedonia group. Additional analysis of which part of the covariance were important for classification, revealed that the only feature surviving correction for multiple comparisons was the variance within the left TPJ. This region has been associated with social cognition and ToM in several previous studies (Bodnar et al., 2014; Dodell-Feder, Tully, Lincoln, & Hooker, 2014; Kronbichler, Tschernegg, Martin, Schurz, & Kronbichler, 2017) and was also a prominent region in the decomposition methods to presented below. For more details on this analysis wee the "interpretation of covariance features for HSA classification" section in Data S1. The second highest classification performance was obtained when using the time series from the inferior lateral prefrontal cortex seed (MCC = 0.35, $p_{un\text{-}corrected}$ = .007), however, classification was not significant after multiple comparisons correction, which was necessary since classification was performed for each seed region separately.

On the contrary, several features from the *decomposition methods* yielded significant classification even after multiple comparisons correction. Here, we used both the time series and spatial maps for each network as classification feature, and corrected for multiple comparisons using maximum permutation statistics across components. For each decomposition method, only one (or sometimes no) feature yielded significant classification.

The highest classification performance was obtained when using one time series from the wbMSAA approach (MCC = 0.56, p = .008). Very interestingly this was the same TS that also obtained the highest task classification performance for the ToM condition, highlighting the coupling between schizotypy and ToM processing (Bora & Pantelis, 2013; Pickup, 2006). In future studies, such coupling between schizotypy and a relevant task (e.g., ToM), could be used to preselect relevant network features instead of testing the classification for all features extracted by the MSAA.

Furthermore, the spatial map corresponding to this time series also obtained the highest classification performance of all wbMSAA spatial maps, which was borderline significant (MCC = 0.42, p = .09, $p_{un\text{-}corrected}$ = .006). The seed of this network was in the TPJ, and the network further included inferior and medial PFC and insula.

wbMSAA

Back view

sMSAA_Lit

Back view

ICA

Back view

sMSAA_Pcon

Back view
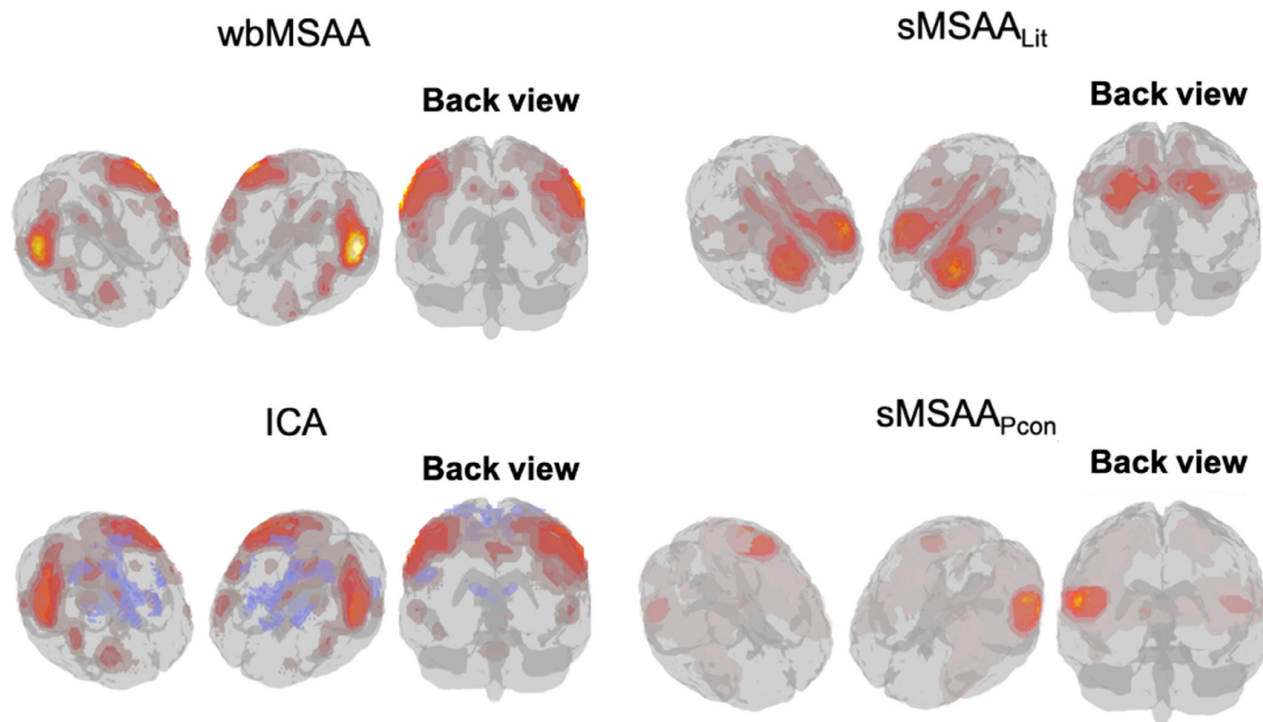
FIGURE 6 Mean spatial maps of the components yielding significant HSA classification performance. For all features, highest classification performance was obtained when using the times series (TS). This figure shows the corresponding spatial maps. The top row shows the two networks, corresponding to the TS features that obtained significant classification after multiple comparisons correction (wbMSAA and SMSAA_Lit). Visualization threshold was 10% fractional contribution. The networks in the bottom row are from the ICA and sMSAAPCon analysis, where the un-corrected p-value was below .05. For visualization, the ICA map was thresholded at a Z-score of 1 and the sMSAA_PCon network include voxels with 5% fractional contribution [Color figure can be viewed at wileyonlinelibrary.com]

Furthermore, the ICA feature (time series) that resulted in the highest classification performance (MCC = 0.45, p = .07, $p_{uncorrected}$ = .005), had a corresponding spatial map, that was nearly identical to the network from the wbMSAA analysis (see Figure 6 and Data S1–Figure S5).

The second decomposition feature that yielded significant classification, was the time series from the spotlight sMSAA_lit approach (MCC = 0.49, p = .03). The spatial map corresponding to this TS had its seed region is in the dorsolateral PFC and furthermore the network included cingulate cortex and motor areas. For the sMSAA_lit approach, we chose the seed regions which were known to be involved in ToM and empathy processing, since it is well established that social cognition is reduced in patients with schizophrenia (Bora et al., 2009; Brunet et al., 2000), and in subjects with schizotypy (Ettinger et al., 2015; Pickup, 2006).

As described in Section 2.8, we also tested another spotlight approach where we used the peak coordinates from a pooled condition analysis (sMSAA_PCon), because this would be a more data driven way to choose center coordinates. Since the pooled condition analysis was specific for the given task, we hypothesized that the features extracted by this approach would result in a higher classification performance than for the sMSAA_Lit approach. However, neither of the time series or spatial maps from the sMSAA_PCon analysis resulted in significant classification after multiple comparisons correction. Only one component (time series) obtained a classification performance, with an un-corrected p-value <.05. Most activation in this network was in the TPJ and angular gyrus, but also included thalamus, insula, and i/m FG.

To summarize, the components from the decomposition methods, which obtained the highest classification performance generally included temporoparietal and prefrontal regions, as well as insula and cingulate cortex. These findings are in accordance with earlier studies which have reported lower white matter integrity in the fronto-temporal tracts (measured by diffusion tensor imaging) in subjects with a high degree of schizotypy (Nelson et al., 2011), and both structural as well as functional studies have related changes in the PFC to schizotypy (Kühn et al., 2012; Raine, Sheard, Reynolds, & Lencz, 1992). Furthermore, earlier studies have shown a decrease in insula gray matter volume in UHR groups (Chan et al., 2011), and it has even been suggested that structural insular abnormalities might be related to the vulnerability for the development of later psychosis (Takahashi et al., 2009). In future studies, it could thus be interesting to investigate if functional imaging could support the structural findings of Takahasi et al., and maybe enable identification of schizotypy in even earlier stages than what is possible with the structural changes. As for insula, gray matter volume reductions in thalamus have also been found in both schizophrenia (Ettinger et al., 2001) as well as in schizotypy (Kühn et al., 2012). Furthermore, fMRI studies have shown

correlation between reduced activation in thalamus and the degree of schizotypy (Aichert, Williams, Möller, Kumari, & Ettinger, 2012; Kumari, Antonova, & Geyer, 2008), but it should be noted that the subjects performed another task in these studies.

In summary, the included areas of the two networks which are able to significantly classify HSA, have consistently been related to schizotypy and the schizophrenia development, which highlight the potential importance of these networks.

Finally, we want to comment on the use of *spatial and temporal network features* for the classification. Whereas many spatial network features resulted in significant classification of the task conditions, the time series generally resulted in a higher classification performance for the social anhedonia classification. This finding indicates that the temporal dynamics during the social cognition task captures important information to differentiate between high and low social anhedonia. In comparison, the connectivity measures used to extract spatial network features in this study are regarded static. In future studies, it would thus be interesting to look at dynamic functional connectivity, where the connectivity is estimated repeatedly for different windows of the time series, and thus also reflect the dynamic variations in the time series (Hutchison et al., 2013) (Damaraju et al., 2014; Nielsen et al., 2018).

3.4 | Discussion of the MSAA method

This study is one of the first to use the MSAA method on neuroimaging data, and the first to implement the spotlight approach that further bridges aspects of data-driven decomposition methods and seed based analysis. We, therefore, highlight some of the important aspects of MSAA.

3.4.1 | Interpretability

Due to the nonnegativity and sum-to-one constraints, the spatial maps in MSAA have a clear interpretation, showing the fractional contributions of the components (archetypes) at each voxel. We used a threshold of 0.1 for visualization, meaning that for each voxel shown in a spatial map, this component had a relative contribution of at least 10% to that given observation. A similar interpretation of the scale in ICA is not immediately possible without additional post processing, and furthermore as the ICA allows both positive and negative contributions, the components can include cancelation effects leading to less straightforward interpretation.

3.4.2 | Noise modeling

The MSAA approach enables heteroscedastic noise modeling, that is, the noise can be estimated for each subject and each voxel separately, instead of assuming it to be constant, which is done in previous decomposition methods such as ICA. Visual inspection of the spatial distribution of these noise levels (Data S1–Figure S3) showed that most noise was present around the edges of the brain and close to known major blood vessels, which probably reflects residual

movement effects and noise due to blood pulsation, respectively. A more elaborate discussion of this noise modeling can be found in (Hinrich et al., 2016).

3.4.3 | Spotlight

The spotlight restriction of MSAA showed to successfully enforce the algorithm to reveal functional networks, which otherwise were obscured by other salient signal features. This is somewhat similar to what was done by seed based analysis, but for the spotlight MSAA the optimal seed is determined by data driven optimization instead of manual assignment. Restriction of the seed regions can be especially valuable if a specific hypothesis needs to be tested, for example, how the connectivity between the whole brain and a particular region changes in relation to disease progression. However, compared to the wbMSAA approach, it requires the user to choose a number of seed regions, which can be difficult to choose. In this study, we have chosen center coordinates based on the social cognition task, either based on previous literature or from a pooled condition analysis. Another approach could have been to choose seeds, which have been related to social anhedonia and/or schizotypy progression.

3.4.4 | Nonconvex optimization and number of components

As for ICA, the MSAA algorithm is a nonconvex optimization problem, which means that the optimization might get struck in a local and not global minimum. In practice, this means that repeated runs can result in somewhat different networks. How severe this problem is, depends on the stability of the given dataset (signal to noise ratio, intersubject differences, etc.) as well as on the number of components chosen. In this study, we used 25 components as this was found to be the optimal number using the minimum description length criteria, which is the default implemented in GIFT toolbox (Li et al., 2007). Using 25 components, resulted in relatively stable networks, with a mean spatial correlation of 86% for the wbMSAA when choosing the best (lowest cost) solution between 10 runs as described in Section 2.7. Visual inspection of the networks showed that the overall network structures between runs were very stable, and the nonperfect machining resulted in small network differences between runs (networks illustrated in Data S1 section "stability of wbMSAA"). Furthermore, we noted that the number of components within one run seemed reasonable, such that known networks were captured by separate spatial maps (mixing of e.g., task and visual processing networks would indicate that the number of components was too low), and did not split networks up into separate components (this would indicate that the number of components was too high). All in all, this indicates that the number of runs and components were appropriate for the given study. However, we want to emphasize the importance of investigating the stability in future studies applying MSAA.

3.4.5 | Toolbox

We have implemented the MSAA (both whole brain and spotlight) code into a SPM plugin (compatible with SPM 12), which interested users can download here: http://www.brain-fmri.com/MSAA/. The plugin enables the user to apply the MSAA algorithm on fMRI data, by simply loading the preprocessed images and choose the optimization parameters specified in the toolbox.

3.5 | Limitations and future perspectives

As discussed in the previous section there are some challenges for decomposition methods, such as nonconvexity and choosing an appropriate number of components. Another large challenge of this study was the relatively small difference between subjects of the high and low social anhedonia respectively. Firstly, classification was challenged by the low separation boundary which was used (mean plus one standard deviation). Though similar boundaries have been used in previous group comparison studies of schizotypy (Wang et al., 2016), it was challenging for the support vector machine to learn from the data of two relatively similar classes. Secondly, even with this low separation threshold, we had an unbalanced dataset, with 56 (LSA) and 14 (HSA) subjects in each group. This further challenged the supervised classification procedure, and made the classification performance sensitive the classification of few subjects. We tried to mitigate this problem by (a) using weights in the support vector machine to counteract the imbalance and (b) used the MCC measure to access classification performance. Additionally, it is important to note that while full correction of multiple comparisons was considered within each feature extraction method, this was not done across these different methods. This was motivated by the main aim of comparing a set of, in many aspects, very similar feature extraction methods. With these limitations in mind, we consider the present study an explorative investigation of features for classification of social anhedonia rather than a study of the neural correlates of social anhedonia itself. Still, we strongly expect that a larger group, particularly with more subjects with high social anhedonia, would make classification easier and more stable. Furthermore, including subjects with more pronounced social anhedonia, or subjects belonging to other risk groups, would also be very interesting from a clinical perspective.

However, even with these challenges, the whole brain and spotlight MSAA algorithms extracted features that yielded significant classification. Using the same methods on ultra-high-risk groups or patients with schizophrenia would thus be very interesting to investigate how networks alterations are related to the development of schizophrenia. Optimally, this could be investigated through a longitudinal study starting with a large group of subjects with a continuous range of schizotypy and a specific and well-designed experimental set-up.

4 | CONCLUSION

Using a variety of different feature extraction methods, we found significant classification of social anhedonia for two features, both consisting of times series extracted by the MSAA decomposition methods. The highest classification performance was achieved using the whole brain MSAA. Importantly, the same time series also obtained the highest task classification performance, making a strong coupling between the processing of the ToM task and the degree of social anhedonia. This indicates that future studies could focus on components representing task-relevant networks for classification of schizotypy, thereby circumventing the need for correction for multiple comparisons across components. The spatial map corresponding to the time series yielding highest classification performance, included the TPJ, prefrontal cortex, angular gyrus and insula, which all have been consistently related to schizotypy as well as to the development of schizophrenia in earlier studies.

Finally, a nearly identical feature was also identified as the best performing when using features extracted by ICA. The repeated occurrence of the same feature highlights the potential importance of this network for early identification of schizotypy. Thus, in future studies, it would be very interesting to investigate if the same network would also be important for subjects with more pronounced schizotypy and other high-risk groups through the spectrum of schizophrenia development.

CONFLICT OF INTEREST

The authors have no professional or financial interests that could be perceived as having biased the presentation.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding authors. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Laerke Gebser Krohne ⓘD https://orcid.org/0000-0001-5354-1482
Yi Wang ⓘD https://orcid.org/0000-0001-6880-5831
Jesper L. Hinrich ⓘD https://orcid.org/0000-0003-0258-7151
Morten Moerup ⓘD https://orcid.org/0000-0003-4985-4368
Raymond C. K. Chan ⓘD https://orcid.org/0000-0002-3414-450X
Kristoffer H. Madsen ⓘD https://orcid.org/0000-0001-8606-7641

REFERENCES

Abu-Akel, A., & Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia*, 49, 2971–2984.

Aichert, D. S., Williams, S. C. R., Möller, H. J., Kumari, V., & Ettinger, U. (2012). Functional neural correlates of psychometric schizotypy: An fMRI study of antisaccades. *Psychophysiology*, 49, 345–356.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews. Neuroscience*, 7, 268–277.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron*, 65, 550–562.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. a., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16, 412–424.

Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23, 137–152.

Bedwell, J. S., Compton, M. T., Jentsch, F. G., Deptula, A. E., Goulding, S. M., & Tone, E. B. (2014). Latent factor modeling of four Schizotypy dimensions with theory of mind and empathy. *PLoS One, 9*, e113853.

Benedetti, F., Bernasconi, A., Bosia, M., & Smeraldi, E. (2009). Functional and structural brain correlates of theory of mind and empathy deficits in schizophrenia. *Schizophrenia Research*, 114, 154–160.

Biswal, B., Yetkin, F. Z., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medical Sciences*, 34, 537–541.

Blanchard, J. J., Collins, L. M., Aghevli, M., Leung, W. W., & Cohen, A. S. (2011). Social anhedonia and schizotypy in a community sample: The Maryland longitudinal study of schizotypy. *Schizophrenia Bulletin*, 37, 587–602.

Bodnar, M., Hovington, C. L., Buchy, L., Malla, A. K., Joober, R., & Lepage, M. (2014). Cortical thinning in tempo-parietal junction (TPJ) in non-affective first-episode of psychosis patients with persistent negative symptoms. *PLoS One, 9*, e101372.

Bora, E., & Pantelis, C. (2013). Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: Systematic review and meta-analysis. *Schizophrenia Research*, 144, 31–36.

Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: Meta-analysis. *Schizophrenia Research*, 109, 1–9.

Brunet, E., Sarfati, Y., Hardy-Baylé, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *NeuroImage*, 11, 157–166.

Calhoun V, Adali T, Hansen L (2003). *ICA of functional MRI data: An overview*.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14, 140–151.

Chan, R. C. K., Di, X., McAlonan, G. M., & Gong, Q. Y. (2011). Brain anatomical abnormalities in high-risk individuals, first-episode, and chronic schizophrenia: An activation likelihood estimation meta-analysis of illness progression. *Schizophrenia Bulletin*, 37, 177–188.

Chan, R. C. K., song Shi, H., lei Geng, F., hua Liu, W., Yan, C., Wang, Y., & Gooding, D. C. (2015). The Chapman psychosis-proneness scales: Consistency across culture and time. *Psychiatry Research*, 228, 143–149.

Chan, R. C. K., Wang, Y., Yan, C., Zhao, Q., McGrath, J., Hsi, X., & Stone, W. S. (2012). A study of trait anhedonia in non-clinical chinese samples: Evidence from the chapman scales for physical and social anhedonia. *PLoS One, 7*, 3–8.

Chang, C.-C., & Lin, C.-J. (2011). Libsvm a library for support vector machines. *ACM Trans Intell Syst Technol, 2*, 1–27.

Chapman, L., Chapman, J., Kwapil, T., Eckblad, M., & Zinser, M. (1994). Putatively psychosis-prone subjects 10 years later. *Journal of Abnormal Psychology*, 1689, 171–183.

Cole, D. M., Smith, S. M., & Beckmann, C. F. (2010). Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. *Frontiers in Systems Neuroscience, 4*, 8.

Cortes, C., & Vapnik, V. (1995). *Support-vector networks* (Vol. 297, pp. 273–297). Boston, MA: Academic Publishers.

Cutler, A., & Breiman, L. (1994). Archetypal analysis. *Technometrics, 36*, 338–347.

Damaraju, E., Allen, E. A., Belger, A., Ford, J. M., McEwen, S., Mathalon, D. H., … Calhoun, V. D. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage Clin, 5*, 298–308.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31, 968–980.

Dodell-Feder, D., Tully, L. M., Lincoln, S. H., & Hooker, C. I. (2014). The neural basis of theory of mind and its relationship to social functioning and social anhedonia in individuals with schizophrenia. *NeuroImage Clin, 4*, 154–163.

Erhardt, E. B., Rachakonda, S., Bedrick, E., Allen, E., Adali, T., & Calhoun, V. D. (2011). Comparison of multi-subject ICA methods for analysis of fMRI data. *Brain, 32*, 2075–2095.

Ettinger, U., Chitnis, X. A., Kumari, V., Fannon, D. G., Sumich, A. L., O'Ceallaigh, S., … Sharma, T. (2001). Magnetic resonance imaging of the thalamus in first-episode psychosis. *The American Journal of Psychiatry*, 158, 116–118.

Ettinger, U., Mohr, C., Gooding, D. C., Cohen, A. S., Rapp, A., Haenschel, C., & Park, S. (2015). Cognition and brain function in schizotypy: A selective review. *Schizophrenia Bulletin*, 41, S417–S426.

Fan, Y., Duncan, N. W., de Greck, M., & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience and Biobehavioral Reviews*, 35, 903–911.

Fett, A. K. J., Viechtbauer, W., Dominguez M de, G., & Krabbendam, L. (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis. *Neuroscience and Biobehavioral Reviews*, 35, 573–588.

Fisher, R. A., & Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507–521.

Friston, K. J., Williams, S., Howard, R., & Frackowiak, R. S. J. (1996). Movement related effects in FMRI. *Magnetic Resonance in Medicine, 3*, 346–355.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*, 531–534.

Gooding, D. C., Tallent, K. A., & Matts, C. W. (2005). Clinical status of at-risk individuals 5 years later: Further validation of the psychometric high-risk strategy. *Journal of Abnormal Psychology*, 114, 170–175.

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews. Neuroscience*, 16, 620–631.

Henry, J. D., Bailey, P. E., & Rendell, P. G. (2008). Empathy, social functioning and schizotypy. *Psychiatry Research*, 160, 15–22.

Hinrich, J. L., Bardenfleth, S., Roge, R., Churchill, N., Madsen, K. H., & Morup, M. (2016). Archetypal analysis for modeling multi-subject fMRI data. *IEEE Journal on Selected Topics in Signal Processing, 10*, 1160–1171.

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., … Chang, C. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage, 80*, 360–378.

Insel, T. R. (2010). Rethinking schizophrenia. *Nature, 468*, 187–193.

Kronbichler, L., Tschernegg, M., Martin, A. I., Schurz, M., & Kronbichler, M. (2017). Abnormal brain activation during theory of mind tasks in schizophrenia: A meta-analysis. *Schizophrenia Bulletin*, 43, 1240–1250.

Kühn, S., Schubert, F., & Gallinat, J. (2012). Higher prefrontal cortical thickness in high schizotypal personality trait. *Journal of Psychiatric Research*, 46, 960–965.

Kumari, V., Antonova, E., & Geyer, M. A. (2008). Prepulse inhibition and "psychosis-proneness" in healthy individuals: An fMRI study. *European Psychiatry*, *23*, 274–280.

Kwapil, T. R. (1998). Social anhedonia as a predictor of the development of schizophrenia-spectrum disorders. *Journal of Abnormal Psychology*, *107*, 558–565.

Lagioia, A., Van De Ville, D., Debbané, M., Lazeyras, F., & Eliez, S. (2010). Adolescent resting state networks and their associations with schizotypal trait expression. *Frontiers in Systems Neuroscience*, *4*, 1–12.

Lewis, D. A., & Levitt, P. (2002). Schizophrenia as a disorder of neuro-development. *Annual Review of Neuroscience*, *25*, 409–432.

Li, Y.-O., Adali, T., & Calhoun, V. D. (2007). Estimating the number of independent components for functional magnetic resonance imaging data. *Human Brain Mapping*, *28*, 1251–1266.

Madsen, K. H., Krohne, L. G., Cai, X., Wang, Y., & Chan, R. C. K. (2018). Perspectives on machine learning for classification of Schizotypy using fMRI data. *Schizophrenia Bulletin*, *44*, s480–s490.

Mason, O. J. (2015). The assessment of schizotypy and its clinical relevance. *Schizophrenia Bulletin*, *41*, S374–S385.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J., & Sejnowski, T. J. (1998). Analysis of fMRI data by blind separation into independent components. *Human Brain Mapping*, *6*, 160–188.

Modinos, G., Pettersson-Yeo, W., Allen, P., McGuire, P. K., Aleman, A., & Mechelli, A. (2012). Multivariate pattern classification reveals differential brain activation during emotional processing in individuals with psychosis proneness. *NeuroImage*, *59*, 3033–3041.

Morrison, S. C., Brown, L. A., & Cohen, A. S. (2013). A multidimensional assessment of social cognition in psychometrically defined schizotypy. *Psychiatry Research*, *210*, 1014–1019.

Mørup, M., & Hansen, L. K. (2012). Archetypal analysis for machine learning and data mining. *Neurocomputing*, *80*, 54–63.

Nelson, M. T., Seal, M. L., Pantelis, C., & Phillips, L. J. (2013). Evidence of a dimensional relationship between schizotypy and schizophrenia: A systematic review. *Neuroscience and Biobehavioral Reviews*, *37*, 317–327.

Nelson, M. T., Seal, M. L., Phillips, L. J., Merritt, A. H., Wilson, R., & Pantelis, C. (2011). An investigation of the relationship between cortical connectivity and schizotypy in the general population. *The Journal of Nervous and Mental Disease*, *199*, 348–353.

Nichols, T., & Holmes, A. (2003). Nonparametric permutation tests for functional neuroimaging. *Human Brain Function*, *15*, 887–910.

Nielsen SFV, Levin-Schwartz Y, Vidaurre D, Adali T, Calhoun VD, Madsen KH, Hansen LK, Morup M (2018). *Evaluating models of dynamic functional connectivity using predictive classification accuracy*. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2566–2570.

Patel, A. X., Kundu, P., Rubinov, M., Jones, P. S., Vértes, P. E., Ersche, K. D., ... Bullmore, E. T. (2014). A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. *NeuroImage*, *95*, 287–304.

Penn, D. L., Sanna, L. J., & Roberts, D. L. (2007). Social cognition in schizophrenia: An overview. *Schizophrenia Bulletin*, *34*, 408–411.

Pflum, M. J., & Gooding, D. C. (2018). Context matters: Social cognition task performance in psychometric schizotypes. *Psychiatry Research*, *264*, 398–403.

Pickup, G. J. (2006). Theory of mind and its relation to schizotypy. *Cognitive Neuropsychiatry*, *11*, 177–192.

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, *17*, 1510–1517.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*, 2142–2154.

Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*, 37–63.

Rachakonda S, Egolf E, Correa N, Calhoun V, Neuropsychiatry O (2015). *Group ICA/IVA of fMRI toolbox (GIFT) manual*.

Raine, A., Sheard, C., Reynolds, G. P., & Lencz, T. (1992). Pre-frontal structural and functional deficits associated with individual differences in schizotypal personality. *Schizophrenia Research*, *7*, 237–247.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34.

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S.-J., De Brito, S. A., McCrory, E. J. P., & Viding, E. (2012). Neural processing associated with cognitive and affective theory of mind in adolescents and adults. *Social Cognitive and Affective Neuroscience*, *7*, 53–63.

Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*, *46*, 668–677.

Shinkareva, S. V., Ombao, H. C., Sutton, B. P., Mohanty, A., & Miller, G. A. (2006). Classification of functional brain images with a spatio-temporal dissimilarity map. *NeuroImage*, *33*, 63–71.

Takahashi, T., Wood, S. J., Yung, A. R., Velakoulis, D., & Pantelis, C. (2009). Insular cortex gray matter changes in individuals at ultra-high-risk of developing psychosis. *Schizophrenia Research*, *111*, 94–102.

Takeuchi, H., Taki, Y., Nouchi, R., Sekiguchi, A., Hashizume, H., Sassa, Y., ... Kawashima, R. (2014). Association between resting-state functional connectivity and empathizing/systemizing. *NeuroImage*, *99*, 312–322.

Thakkar, K. N., & Park, S. (2010). Empathy, schizotypy, and visuospatial transformations. *Cognitive Neuropsychiatry*, *15*, 477–500.

Völlm, B. A., Taylor, A. N. W., Richardson, P., Corcoran, R., Stirling, J., McKie, S., ... Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a non-verbal task. *NeuroImage*, *29*, 90–98.

Wang, Y., Li, Z., Liu, W., Wei, X., Jiang, X., Lui, S. S. Y., ... Wei, X. (2018). Negative schizotypy and altered functional connectivity during facial emotion processing. *Schizophrenia Bulletin*, *44*, S491–S500.

Wang, Y., Liu, W.-H., Li, Z., Wei, X.-H., Jiang, X., Neumann, D. L., ... Chan, R. C. K. (2015). Dimensional schizotypy and social cognition: An fMRI imaging study. *Frontiers in Behavioral Neuroscience*, *9*, 133.

Wang, Y., Liu, W. H., Li, Z., Wei, X. H., Jiang, X. Q., Geng, F. L., ... Chan, R. C. (2016). Altered corticostriatal functional connectivity in individuals with high social anhedonia. *Psychological Medicine*, *46*, 125–135.

Wang, Y., Lui, S. S. Y., quan, Z. L., Zhang, Q., Zhao, Q., Yan, C., ... Chan, R. C. K. (2014). Individuals with psychometric schizotypy show similar social but not physical anhedonia to patients with schizophrenia. *Psychiatry Research*, *216*, 161–167.

Wang, Y., Neumann, D. L., Shum, D. H. K., Liu, W., Shi, H., Yan, C., ... Chan, R. C. K. (2013). Cognitive empathy partially mediates the association between negative schizotypy traits and social functioning. *Psychiatry Research*, *210*, 62–68.

Weinberger, D. (1987). Implications of normal brain development for the pathogenesis of schizophrenia. *Archives of General Psychiatry*, *44*, 660–669.

SUPPORTING INFORMATION

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

Supplementary Material

This supplementary material includes additional figures and tables which are referred to in the main paper, and that are described by their corresponding captions. In short, **supplementary table** 1 lists the literature center coordinates that are used for the seed region analysis and spotlight multi-subject archetypal analysis (MSAA).

Supplementary table 2 gives additional information on the Chapman and the Beck depression inventory scales.

Supplementary figure 1-4 shows all networks that obtained significant task classification for either the "Theory of Mind" or "Empathy" condition, furthermore the figure lists the accuracies obtained by each network.

Supplementary figure 5 shows axial slices of the networks that obtained significant classification for the social anhedonia classification.

Supplementary figure 6 illustrates the average (over subjects) noise map for the whole brain MSAA.

The section **stability of wbMSAA** describes the consistency of the wbMSAA analysis across multiple runs of the algorithm.

The section **Interpretation of covariance features for HSA classification** investigate which ROIs were responsible for the significant classification of HSA based on covariance features.

Finally, **supplementary table 3** includes the center coordinates from the pooled condition analysis that were used for the spotlight MSAA classification.

Supplementary table 1: MNI coordinates of the 25 center coordinates used for the seed region analysis and spotlight sMSAA$_{Lit}$. The number in the brackets indicate the network number (right before left), e.g. TPJ (1-2) indicates that rTPJ was seed number 1.

Literature coordinates	Left hemisphere MNI coordinate			Right hemisphere MNI coordinates		
	x	y	z	X	y	Z
Temporoparietal junction (TPJ) (1-2)	-53	-59	20	53	-59	20
Inferior parietal Lobe (IPL) (3-4)	-45	-43	56	45	-43	56
Angular gyrus (AngG) (5-6)	-45	-60	35	45	-60	35
Superior Temporal Sulcus (STS) (7-8)	-54	-12	0	54	-12	0
Precuneus (Pcun) (9)	-2	-52	39			
Amygdala (Amyg) (10-11)	-20	-3	-18	20	-3	-18
Ventral Striatum (12-13)	-10	15	9	10	8	-8
Dorsal Temporal Pole (dTP) (14-15)	-54	-9	-21	54	-9	-21
Dorsal Anterior Cingulate Cortex (dACC) (16-17)	-10	32	24	10	32	24
Ventral Anterior Cingulate Cortex (vACC) (18-19)	-12	28	-10	12	28	-10
Ventral medial Prefrontal Cortex (vmPFC) (20)				3	51	-15
Dorsal medial Prefrontal Cortex (dmPFC) (21)				6	26	55
Dorsolateral Prefrontal Cortex (dLPFC) (22-23)	-33	38	37	33	38	37
Inferiorlateral Prefrontal Cortex (dLPFC) (24-25)	-46	22	8	46	22	8

Supplementary table 2: Chapman scale scores and Beck Depression inventory for all 70 included subjects.

|	Chapman scale scores				Beck Depression Inventory (BDI)
	Chapman Social Anhedonia (CSAS)	Chapman Physical Anhedonia (CPAS)	Magical ideation (MIS)	Perceptual aberration (PAS)	
Mean	7.97	12.59	6.59	10.44	4,06
Standard deviation	5.65	9.87	7.06	5.64	4.51

Supplementary figure 1: Significant task classification networks from SBA analysis. 3D visualization of all networks that obtained significant task classification for either the theory of mind (left and middle column) and empathy (right column) classification using the 25 networks coming from seed based analysis (SBA). The network number (no.) corresponds to the seed region number listed in supplementary table 1.

Seed based analysis

Theory of Mind
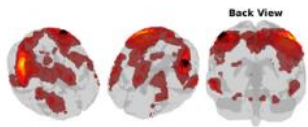
Network no. 4
Accuracy = 71 % p < 0.001
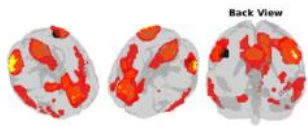
Network no. 5
Accuracy = 71 % p < 0.001

Network no. 9
Accuracy = 71 % p < 0.001
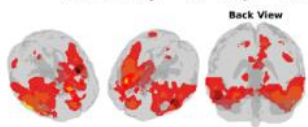
Network no. 15
Accuracy = 71 % p < 0.001

Theory of Mind
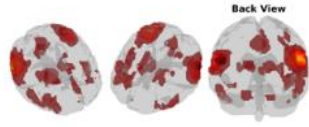
Network no. 1
Accuracy = 67 % p < 0.01
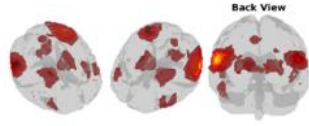
Network no. 2
Accuracy = 67 % p < 0.01
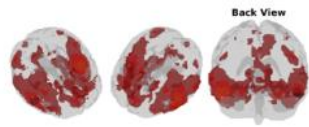
Network no. 14
Accuracy = 67 % p < 0.01
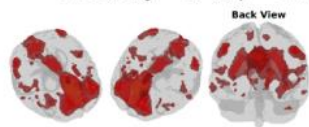
Network no. 17
Accuracy = 67 % p < 0.01
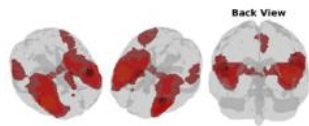
Network no. 25
Accuracy = 66 % p < 0.05

Empathy
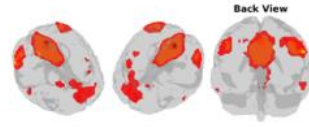
Network no. 9
Accuracy = 72 % p < 0.001
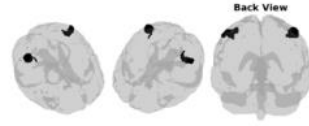
Network no. 3
Accuracy = 70 % p < 0.01

Network no. 4
Accuracy = 69 % p < 0.01
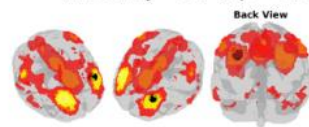
Network no.
Accuracy = 67 % p < 0.01

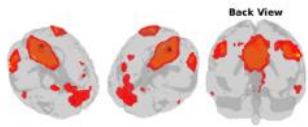**Supplementary figure 2: Significant task classification networks from ICA.** 3D visualization of all networks that obtained significant task classification for either the theory of mind (left column) and empathy (right column) classification using the times series (TS) from the 25 components coming from the independent component analysis (ICA). The component number (no.) corresponds to the order of the networks when returned from the decomposition method, and corresponds to the order of the .nii files available at http://www.brain-fmri.com/MSAA/supplement/.
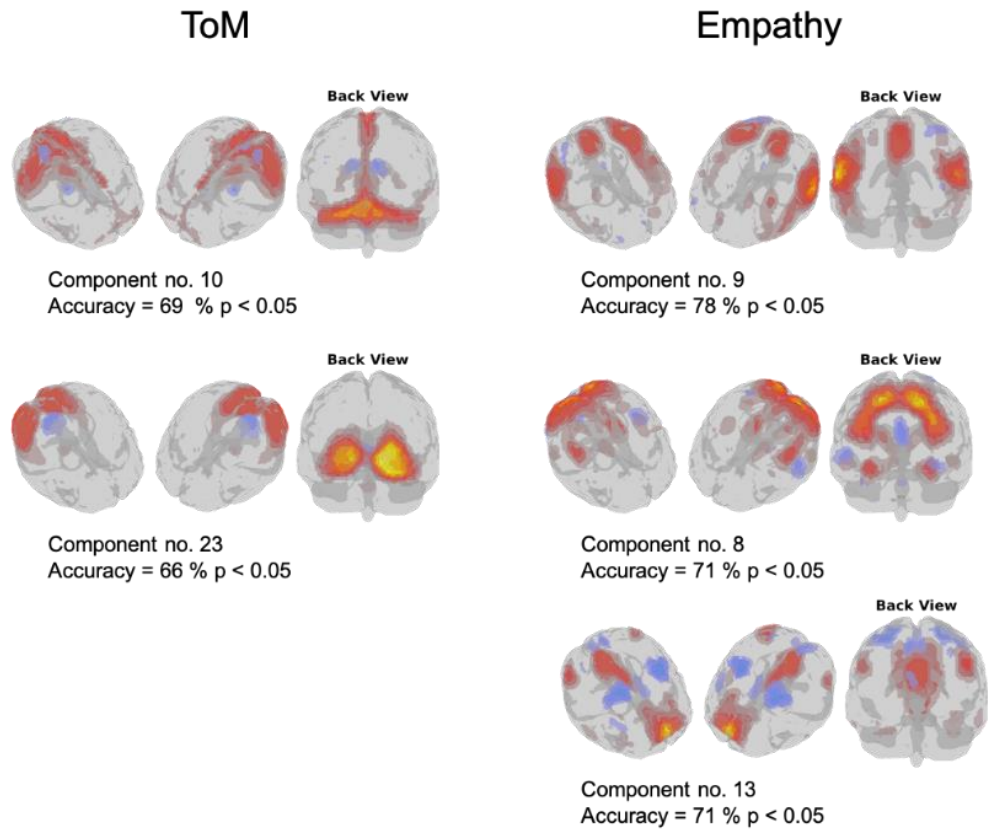
Supplementary figure 3: Significant task classification networks from wbMSAA. 3D visualization of all networks that obtained significant task classification for either the theory of mind (left column) and empathy (right column) classification using the times series (TS) from the 25 components coming whole brain multi subject archetypal analysis (wbMSAA). The component number (no.) corresponds to the order of the networks when returned from the decomposition method, and corresponds to the order of the .nii files available at http://www.brain-fmri.com/MSAA/supplement/.

Supplementary figure 4: Significant task classification networks from sMSAA_Lit. 3D visualization of all networks that obtained significant task classification for either the theory of mind (left column) and empathy (right column) classification using the times series (TS) from the 25 components coming spotlight multi subject archetypal analysis (sMSAA_Lit). The component number (no.) corresponds to the order of the networks when returned from the decomposition method, and corresponds to the order of the .nii files available at http://www.brain-fmri.com/MSAA/supplement/.
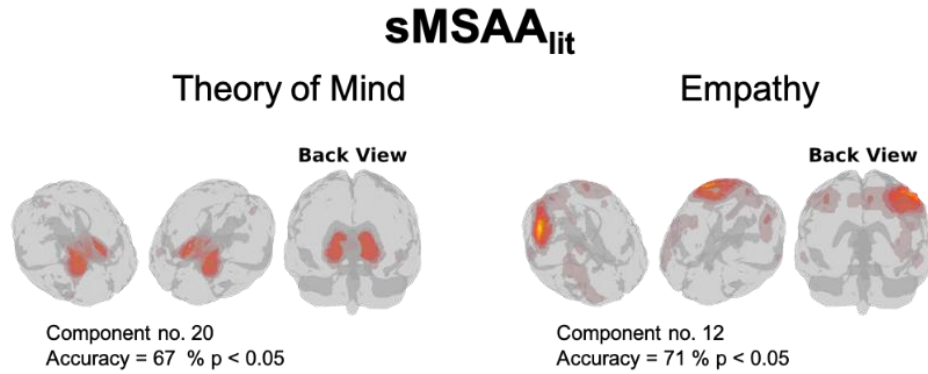
Supplementary figure 5: Axial slices of the best HSA classifying networks determined by the decomposition methods; whole brain MSAA (top left), and spotlight MSAA (bottom) with center coordinates from the literature (sMSAA_lit) (left), and pooled condition analysis (sMSAA_PCon) (right). Finally, axial slices from ICA (top right). For MSAA the visualization threshold was 10% fractional contribution for wbMSAA and sMSAA_Lit and 5% for sMSAA_PCon. For visualization, the ICA map was thresholded at a z-score of 1.
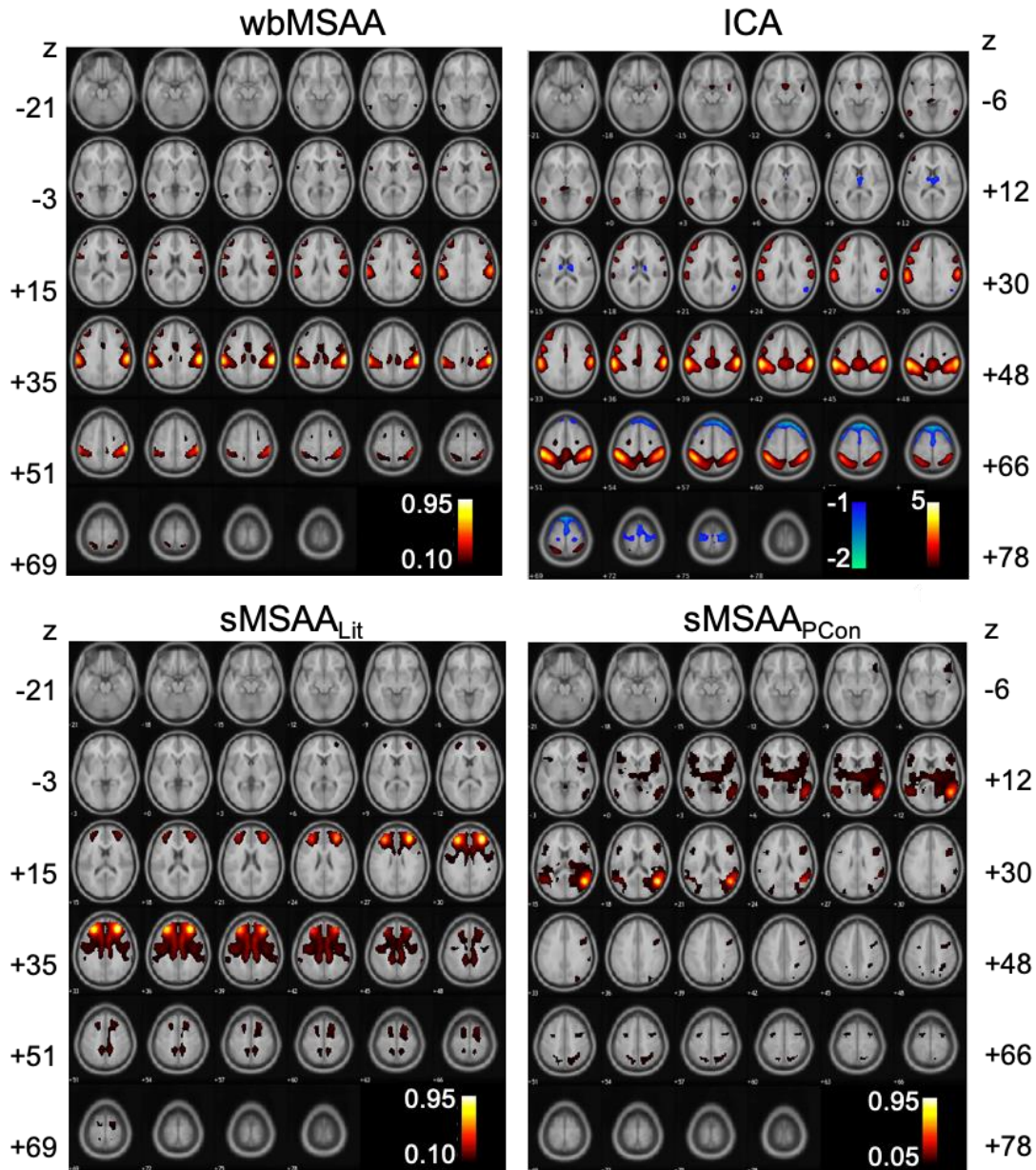
Mean variance across subject (mean noise)
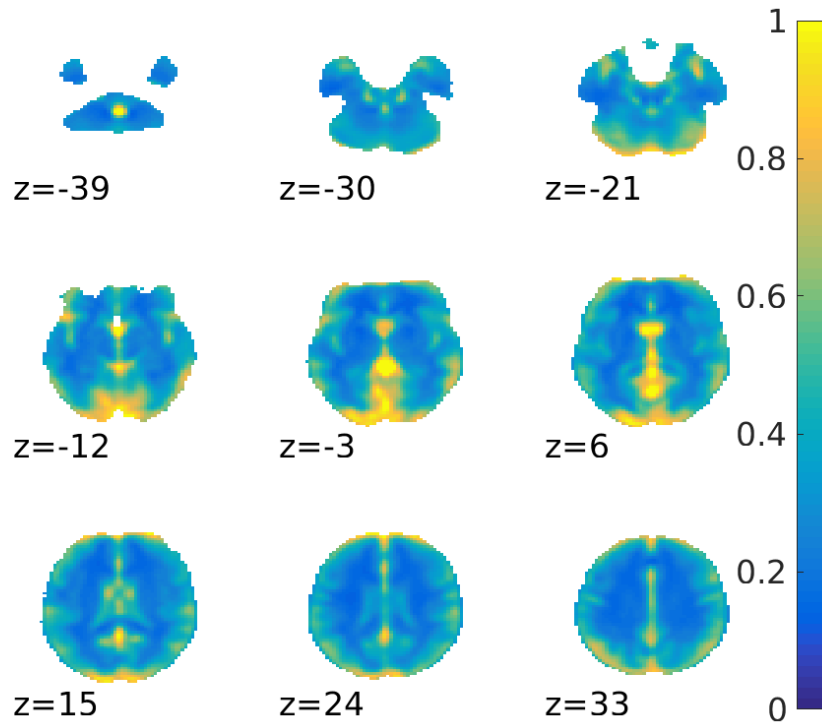
Supplementary figure 6: Mean variance (estimate of noise) between subjects using the wbMSAA algorithm. It is clearly seen that the algorithm determined most noise at the edges and close to big blood vessels, which likely reflect residual movement artifacts and noise due to blood pulsation respectively. Please note that the color scale is arbitrary scaled to 1.

Stability of whole-brain multi-subject archetypical analysis

The cost function in multi-subject archetypical analysis is non-convex just like independent component analysis, hence it is not guaranteed that the same components will be identified across multiple runs with different random initializations. To alleviate this issue, the run which obtained the lowest value of the cost function out of 10 optimizations each with random initializations were chosen each time we used the algorithm. To further investigate the stability, we here rerun this procedure 10 times investigating the similarity of the spatial maps obtained. As there is a trivial ordering/permutation ambiguity of the components across components, they were matched across runs by successively pairing the components that were most correlated (based on correlation of the spatial maps). In this procedure we used the run that was used in the main article as reference, hence the ordering of the components is the same as in the main article.

Supplementary figure 7 shows the spatial correlation of each of the components across the 10 runs. The reference run (run number 9 in supplementary figures 9 and 10) is used as reference and is therefore not shown. For most of the components the components are quite consistent across runs generally obtaining spatial correlations above 0.9, however some components (in particular component 15) are poorly matched across runs. While this indicates some instability, it is not too unexpected in case of model mismatch as some components (in particular unstable nuisance components) may not be identified in all runs. For the components obtaining significant classification rates we generally observe very high correlation across runs indicating high stability, in supplementary figure 8 component number 13 (which obtained significant classification of HSA) is displayed across the 10 runs as an example. Similarly, the least stable component number 15 is displayed in supplementary figure 9. Note, that due to the high dimensionality of the spatial components the correlation value can be low even if the components are visually quite similar.
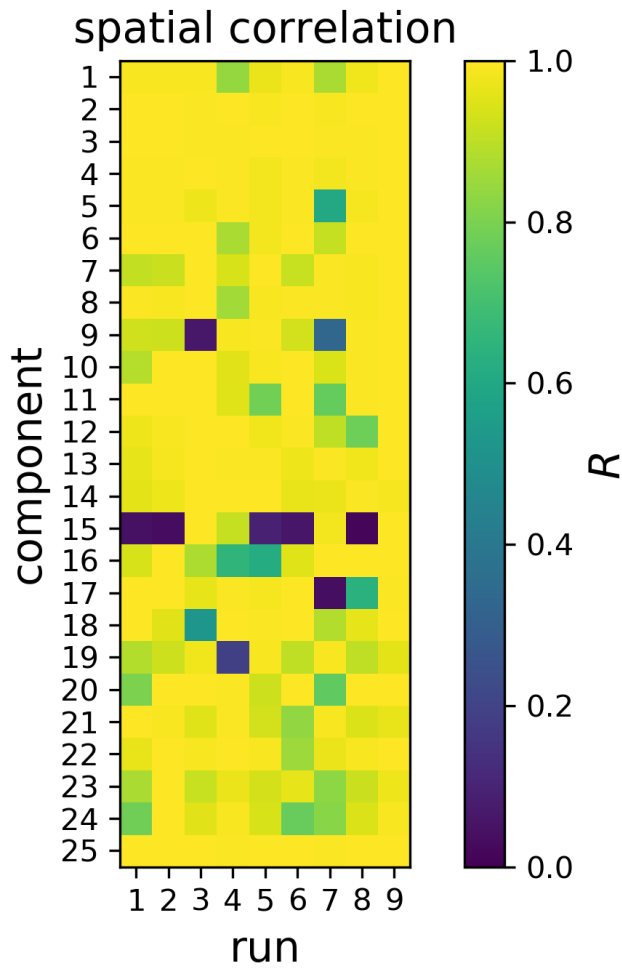
Supplementary figure 7: Correlation of spatial maps across runs and components.
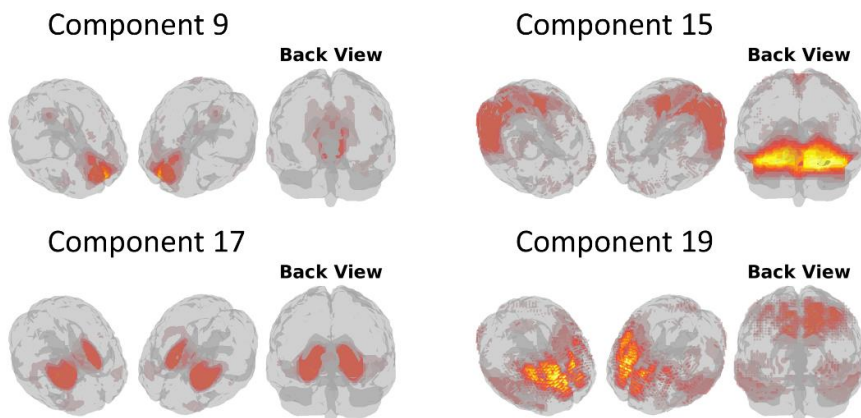
Supplementary figure 8: Spatial maps of the least stable components. The figure shows the spatial maps of the components that were least stable across runs, for brevity only the refence run is shown.
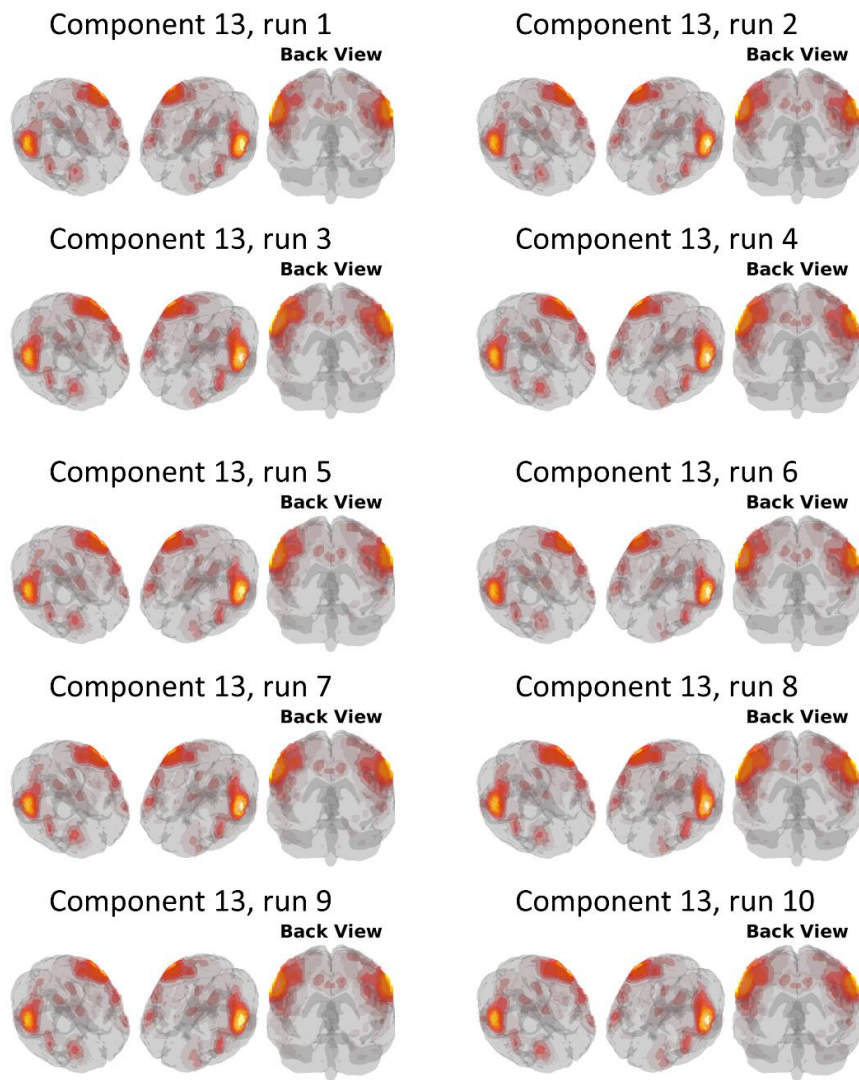
Supplementary figure 9: Spatial maps of component 13. This component obtained significant HSA classification is displayed across the 10 runs. The components are extremely similar across runs, and repeating the HSA classification for each run also showed significant classification in all 10 runs.

Component 15, run 1

Component 15, run 2

Component 15, run 3

Component 15, run 4

Component 13, run 5

Component 15, run 6

Component 15, run 7

Component 15, run 8

Component 15, run 9
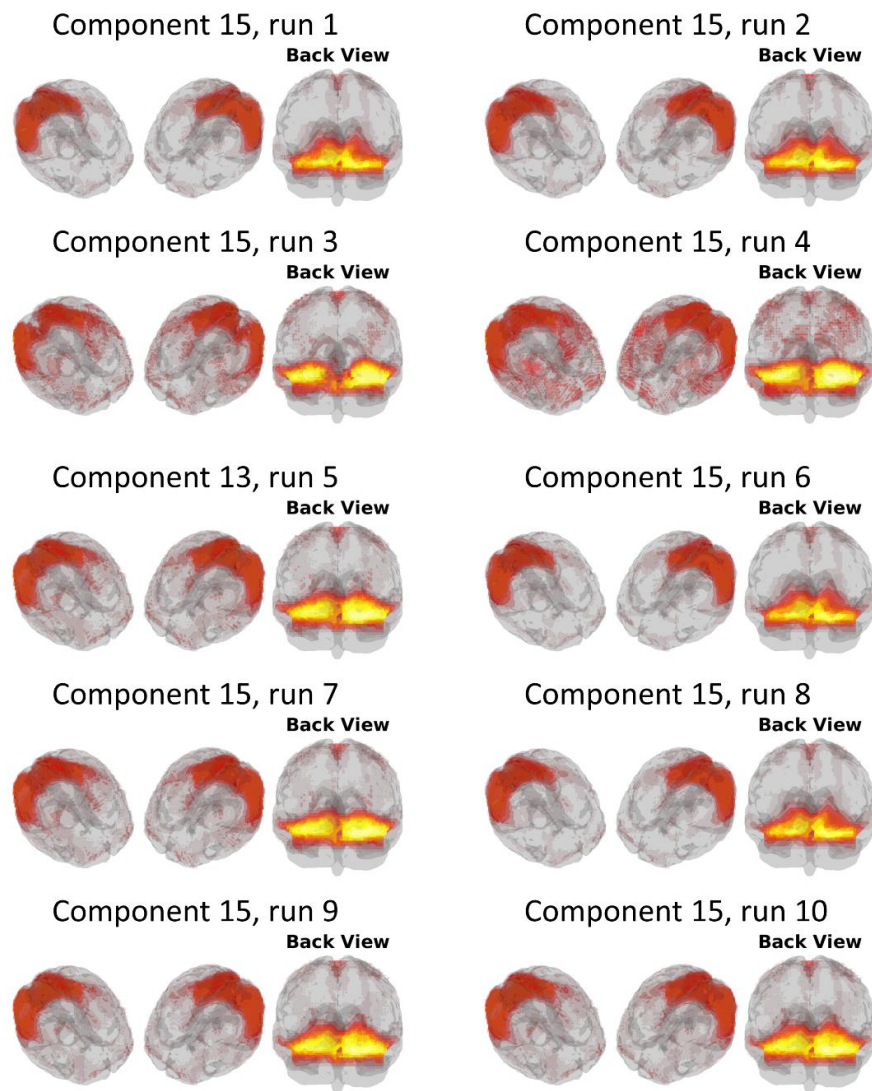
Component 15, run 10

Supplementary figure 10: Spatial maps of component, 15 which was the least stable across runs. Note that despite the low correlation value the components are visually similar across runs.

Interpretation of covariance features for HSA classification

To investigate which of the regions of interest (ROI) and functional connectivity between them were responsible for the significant classification of HSA using covariance features were significant we performed an analysis aiming at reveling the significant features. For this analysis the number of features was the diagonal the upper triangular part of the 25 by 25 ROI covariance matrix resulting in a total of 325 features.

As direct interpretation of weight maps in support vector classification is known to be ambiguous, we used the procedure suggested by Haufe et al. (Haufe et al., 2014) to invert the decoding model to identify an activation map. As statistical inference on activation maps is not immediately possible, we investigated the stability of these maps using repeated cross validation. To this end we used the split-half resampling approach suggested in (Strother et al., 2002) to identify reproducible Z-scored activation maps.

The procedure involved randomly splitting the data into two equally sized proportions (while keeping the proportions of high and low HSA approximately equal in the two proportions). Then the support vector classification where fitted on each of the splits thereby identifying two feature weight vectors w_1 and w_2 in this case with dimensionality 325, these were then converted into activation maps $a_{i,1}$ and $a_{i,2}$ following the procedure suggested in (Haufe et al., 2014)

$$a_i = C_i w_i.$$

where C_i is the estimated 325 by 325 data covariance for split i. Note, that scaling of a_i is arbitrary. Then a reproducible statistical map was constructed as

$$Z = \frac{a_1 + a_2}{\sigma(a_1 - a_2)},$$

Where σ is the standard deviation operator (here subtracting the mean of $a_1 - a_2$ is allowed as for each split as an equivalent opposite split exists). The main insight behind this equation is that $a_1 + a_2$ is an estimate of the activation map while $a_1 - a_2$ is an unbiased estimate of variability of the activation map (due to the independent splits) see (Strother et al., 2002) and (Rasmussen, Hansen, Madsen, Churchill, & Strother, 2012) for further details. In this case we are assuming equal variance across features by using a scalar covariance estimate. The result is an approximately z-scored (under an assumption of normality) activation map. The estimate can be improved by averaging across repeated splits, in this study we averaged across 100 splits of the data. To do inference on the significant features we compared the averaged z-scored map to a cumulative Normal distribution and Bonferroni corrected across the 325 multiple comparisons to control two-sided family-wise type I error at the 5% level leading to a Z threshold of $\Phi^{-1}\left(\frac{0.05/2}{325}\right) \approx 3.78$, where Φ^{-1} denotes the inverse normal cumulative distribution function. Supplementary figure 11 shows the significant features. Note that only feature surviving the family-wise error correction is the variance within the left temporal parietal junction.
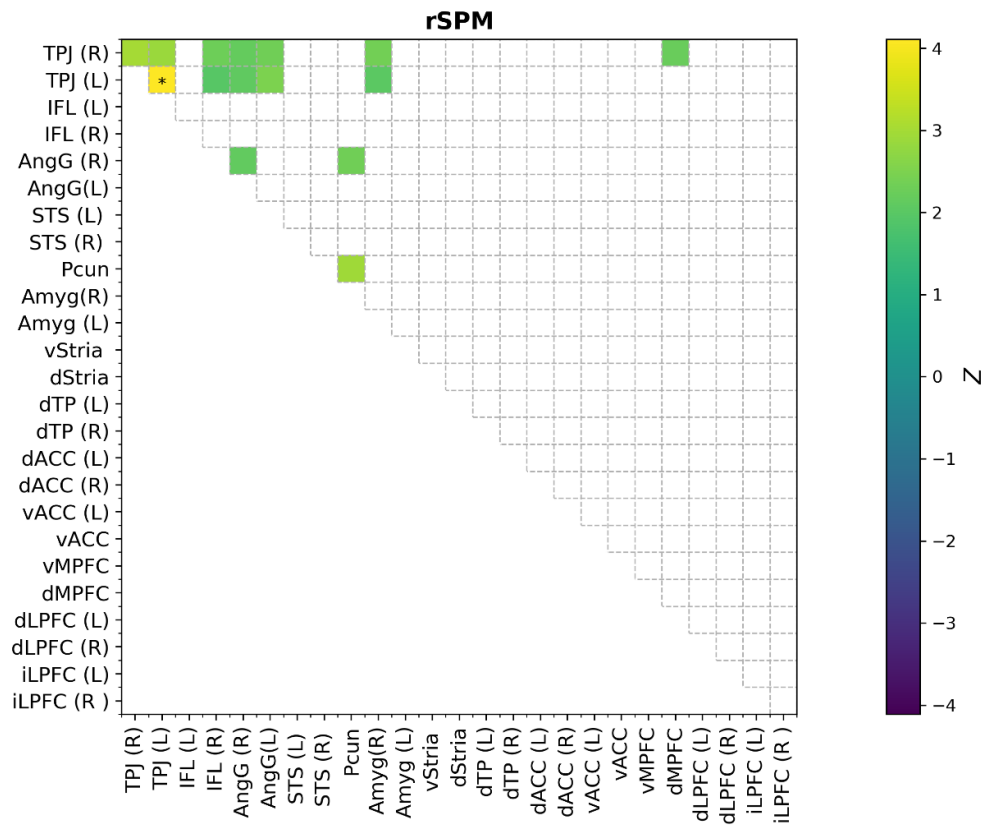
Supplementary figure 11: Reproducible statistical map of covariance features for classification of HSA, z-scores significant at the two-sided 5% significance level is shown and significance at 5% Bonferroni corrected for multiple comparisions is indicated by a *. Only the covariance of the left temporal parietal junction, TPJ (L) reach significance.

Supplementary table 3: Results from pooled condition (PCon) mass univariate analysis. Center coordinates from this analysis were used for the PCon spotlight mask used for MSAA. Table shows peak Z score, cluster size, and MNI coordinates for all significant clusters (Significance level $\alpha_{RFT}\leq0.05$, where random field theory was used to correct for multiple comparison correction).

	Left hemisphere					Right hemisphere				
	Peak Z	Cluster size	MNI coordinate x	y	z	Peak Z	Cluster size	MNI coordinates x	y	z
Pooled condition analysis (PCon)										
Anterior middle temporal gyrus	6.15	109	-54	-9	-21	7.28	235	54	0	-21
Temporoparietal junction	7.30	287	-51	-69	24	7.12	332	51	-63	24
Lateral Superior temporal sulcus	4.75	6	-42	18	-33					
Medal superior temporal sulcus	4.62	1	-36	21	-33					
Cuneus	7.56	135	-12	-108	9	7.23	93	15	-105	12
Ventral medial prefrontal cortex						5	19	3	51	-15
Dorsal medial frontal gyrus	5.63	46	-12	51	48	4.7	3	12	57	42
Anterior parahippocampal gyrus	5.38	30	-24	15	-21					
Posterior parahippocampal gyrus	4.71	8	-21	-33	-18					
Fusiform gyrus	5.05	7	-36	-48	-21					
Middle occipital gyrus	4.81	7	-45	-84	0					
Precuneus	7.37	603	0	-57	39					

References

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W., & Strother, S. C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, *45*(6), 2085–2100. https://doi.org/10.1016/j.patcog.2011.09.011

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., … Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *NeuroImage*, *15*(4), 747–771. https://doi.org/10.1006/nimg.2001.1034

PAPER C

D

Title

Neuroimaging based predictions of Schizophrenia diagnosis and PANSS scores, a multi-site resting state fMRI study

Authors

Krohne, Laerke G ;Christensen, Søren R; Moerup, Morten; Madsen, Kristoffer H

Status

In preparation

Title: Neuroimaging based predictions of Schizophrenia diagnosis and PANSS scores, a multi-site resting state fMRI study

Authors:

Lærke G Krohne [1,2], Søren R. Christensen [2], Morten Mørup [1], Kristoffer H. Madsen [1,3]

Affiliations

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Experimental Medicine, H. Lundbeck A/S, Valby Denmark

[3] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital - Amager and Hvidovre, Copenhagen, Denmark

Abstract:

Schizophrenia is a complex psychiatric disorder with a high degree of psychopathological heterogeneity and currently there are no clinically used biomarkers to assist diagnostic or treatment decisions. In the search for robust neuroimaging biomarkers, we here combined a large multi-site resting state fMRI (rsfMRI) dataset and machine learning methods for feature extraction and clinical predictions. We used resting state fMRI connectivity features to predict both the diagnostic labels and symptom scores defined using the Positive and Negative Syndrome Scale (PANSS). We merged data from several publicly available databases to obtain data from ten different sites, which we split into a discovery dataset (143 patients with schizophrenia (SZ) and 486 healthy controls (HC) from eight sites) and a validation test dataset (63 SZ patients and 260 HC, from two independent sites). We focused on connectivity changes within and between resting state networks (RSN), which we assessed using three methods: parcellation based connectivity analysis and two decomposition methods: independent component analysis (ICA) and multi-subject archetypal analysis (MSAA). We then performed three prediction analysis, i) classifying patients with SZ from HC, ii) predicting the symptom severity using the total PANSS score, and iii) predicting the positive, negative, and generalized PANSS subscales to address the internal heterogeneity of schizophrenia. For the decomposition methods, we investigated three different transfer learning approaches to determine ways to bridge decomposition features across datasets.

For the feature extraction analyses, we found that both decomposition methods extracted 14 similar RSNs that were stable across datasets. The diagnosis classification was high and significant on both the discovery and validation dataset for all three feature extraction methods, and the highest performance was found using classifiers that included data from all RSNs. On the contrary, the prediction of the PANSS scores were only low to moderate and overall, the models generalized poorly to the validation dataset. We see the outcomes of this study as important methodological contributions towards using machine learning and multi-site imaging for predictive modelling. We hope that data sharing initiatives will continue and expand, as we believe that even more multi-site data from patients, including information about confounding factors, are needed to make firm conclusion on whether machine learning and rsfMRI is the right path to find robust biomarkers to guide clinical decisions for patients with schizophrenia.

Highlights

- ICA and MSAA extract similar and stable (across datasets) RSN from multi-site data
- High and reproducible diagnosis classification with all three methods
- Using transfer learning between the datasets increased the stability and predictive performance
- Best performance with ensemble classifiers indicating no "single best" RSN
- Prediction of PANSS was low to moderate and did not generalize to new data

Keywords:

Clinical predictions, Multi-site fMRI, Resting state connectivity, Schizophrenia, Machine Learning

Introduction

Schizophrenia is a psychiatric syndrome with a complex and heterogeneous neurobiological, genetic, and phenotypic profile. It is usually diagnosed based on the symptomatology using diagnostic tools such as the Diagnostic Manual of Mental Disorders (DSM) manual [1] in the US, or the International Classification of Diseases (ICD)[2] in Europe. The symptom severity is often assessed using the Positive and Negative Syndrome Scale (PANSS), which includes thirty items that are organized into a positive, negative, and general psychopathology subscale [3]. A core challenge of using diagnostic labels and symptom scores, is that they do not necessarily reflect the underlying mechanism that causes them, which means that a set of symptoms can arise from different causes while the same etiology might manifest as different symptoms and phenotypes due to individual differences or environmental factors[4-6] . Unfortunately, there are currently no clinically used objective biomarkers that can inform diagnostic and treatment decisions in schizophrenia (SZ), but neuroimaging is a strong candidate for biomarker discovery [4]. The list of potential neuroimaging biomarkers covers a broad range, from measures of altered release of neurotransmitters (e.g., dopamine), receptor occupancy, neuroinflammation and dysconnectivity between brain regions [4]. In our study, we focused on dysconnectivity biomarkers from multi-site resting state functional magnetic resonance imaging (rsfMRI), i.e., measures of how the functional connectome changes in patients with SZ during rest. Traditionally neuroimaging biomarker discovery has relied on group-level mapping using univariate analytical techniques, but in the last decade an increasing number of studies have used supervised machine learning (ML) to make prediction on an individual level. So far, most neuroimaging predictive modelling studies have focused on classifying the diagnostic labels of participants [5, 6], i.e., using the fMRI data to train a model that can predict if a participant is a SZ patient or a healthy control. Currently there are more than 35 predictive-modelling studies that have used rsfMRI features to classify SZ patients, where most have obtained a high accuracy of 75-90% [7]. However, only few studies have tested their results on data from independent test sites, and for those that have, the predictions performances were substantially lower on the new data[6, 8-11]. Overall, earlier studies (including meta-analyses) have shown widespread functional connectivity changes in patients with schizophrenia, which support the view that it is a disorder of disorganized communication across brain networks[4, 12-15]. However, since the methods and results were variable across studies, firm conclusions are yet to be made [4, 16].

There are some core challenges that have hampered the development of robust neuroimaging biomarkers in psychiatric disorders. These can be grouped into i) internal heterogeneity, ii) technical choices, and iii) clinical utility. For a complete discussion, we refer the reader to previous review

papers, such as Kraguljac et al. from 2021[4-6], but here we introduce some of the main points, and how we aimed to mitigate them in our study.

The **internal heterogeneity** of schizophrenia is a challenge since the current diagnosis only focusses on symptoms, which overlap with many other disorders, and patients with the same diagnosis can be affected by very different symptoms domains. This means that the clinical populations between studies often are different (particularly in studies with small sample sizes) which in turn hampers the generalizability between studies. Several initiatives have been started to find data-driven mechanistic disease definitions or sub-types which have more homogenous biology [16]. One way to do this, is to search for brain activation patterns that related to more find-grained disease definitions that are provided through clinical assessments. This carries great potential, both for biomarker discovery and many other applications such as the development of personalized treatment plans.

With **technical choices** we here refer to challenges in defining best practices both for data acquisition and analysis. Regarding the former, most earlier studies have used single-site data with a limited sample size and strict inclusion criteria (e.g., only including young Caucasian males with first psychosis without comorbidities). Models trained on such datasets have a risk to be overfitted to the narrow "patient space" represented by the specific study, and thus have a poor generalizability to data from independent sites [9, 17]. Furthermore, it has been shown that acquisition parameters also highly influence the reliability and results of the study [18-21]. Secondly, it has been shown that the results of neuroimaging studies strongly depend on the steps included in the analysis, ranging from choices in preprocessing and feature extraction to statistical analysis. All these steps have a high degree of flexibility, and it has been standard practice to adjust the analysis pipeline to the specific dataset (which again can give rise to overfitting) [8, 22, 23].

With **clinical utility** we refer to the overall applicability and health impact of the biomarker. Here, biomarker discovery can be divided into three steps i) analytical validation to establish that the measurement technique reliably measures the intended outcome (for fMRI this is most frequently done using the intraclass correlation coefficient to measure test-re-test reliability [19-21]), ii) clinical validation to establish that the biomarker can adequately predict or measure the relevant clinical concept, e.g., symptom severity. And finally, iii) show that the inclusion of the biomarker improves the outcome in a clinically meaningful way, when assessing both benefits, patient burden and potential risks[24, 25]. Furthermore, to have a clinically useful biomarker, explainability is important [23, 25] to evaluate what brain features were important for the predictions and thus could serve as potential biomarkers. Currently, many activities are ongoing to move machine learning away from a "black box" towards a more refined understanding; however, so far it is often still challenging to

interpret the outcomes of prediction models. For example, neural networks are often too complex to provide meaningful interpretations, and even for "simpler models" like support vector machines utilized in predictive modelling, interpretation of the classification weights should be approached with care [26, 27].

In our study, we aimed to overcome some of these challenges by combining one of the largest available multi-site rsfMRI datasets with machine learning for both data-driven feature extraction and predictive modelling, while having a high focus on robustness and generalizability.

We opted towards keeping all steps in our analyses as robust and data-driven as possible, and used multi-site data to train our models while also keeping data from two independent sites separate for external validation. We deliberately did not perform any site-specific adjustments in our prediction pipeline, with the goal to search for biomarkers that were sufficiently robust to between site variations.

For the feature extraction, we investigated the stability and performance of unsupervised decomposition methods which can be used to extract brain features in a data-driven way. Most frequently, this is done using the decomposition method independent component analysis (ICA) or clustering approaches[28-30]. Multi-subject archetypal analysis (MSAA) is a low rank matrix factorization method that bridges aspects of decomposition and clustering [31, 32], which has shown promising results on an earlier prediction study that we performed on individuals with schizotypy [33]. In this study, we therefore used both ICA and MSAA to extract resting state networks (RSN). Since there is so far no consensus on how to use decomposition methods to search for brain features across datasets, we have investigated three different transfer learning approaches, to bridge networks across datasets. Finally, we have compared the performance of the decomposition methods with features from a parcellation based connectivity analysis, which earlier multi-site prediction studies have used for feature extraction [10, 11, 34].

Based on the connectivity features, we classified participants according to their diagnosis, i.e., classifying patients with schizophrenia from healthy controls. Since firm conclusions are yet to be made about which (or even if any) specific RSNs could serve as potential biomarker for schizophrenia, we performed a comprehensive analysis of the predictive abilities of individual RSNs both for the decomposition methods and the parcellation based connectivity analysis. To our knowledge, no earlier studies have used decomposition methods on multi-site data to classify patients with schizophrenia, nor investigated how transfer learning and individual RSN prediction can be used to increase the stability and explainability of the results.

We also determine if the RSN features could be used to predict the symptom severity (measured using the total PANSS scale) and the three PANSS subscales (positive, negative and generalized), in an attempt to disentangle the internal heterogeneity of schizophrenia. We only know of few earlier predictive modelling studies that have used fMRI data to predict PANSS scores [35-38] and none of these have used multi-site data, tested their findings on external data, nor used decomposition method in their analysis.

We therefore see this work as an important step toward exploring how data-driven machine learning methods and multi-site datasets can be used to search for robust and reproducible biomarkers.

Materials and Methods

Overall, this study includes five different steps which are illustrated in Figure 1 and described in further details in the following sections.
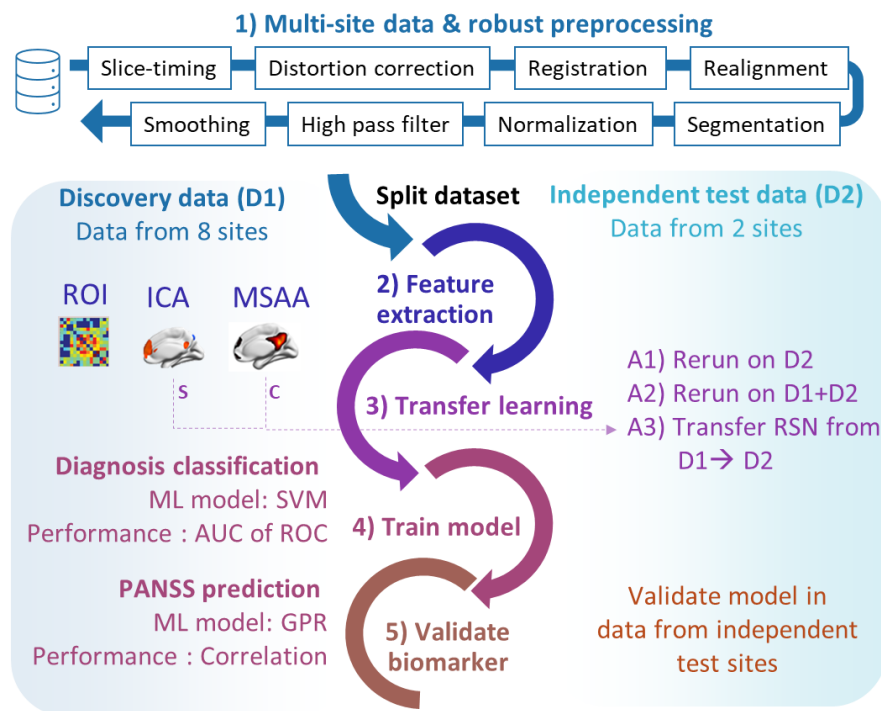
Figure 1 Graphical illustration of the five different steps of this study. 1) The study uses a rsfMRI dataset collected from 10 different sites, which all went through the same robust preprocessing pipeline. The data was split into a discovery dataset (D1) on which the prediction models were trained, and a test dataset (D2) with data from two independent sites, that were used for model validation. 2) Features were extracted using either a parcellation based connectivity (also called region of interest (ROI)) approach, or by use of the data driven decomposition methods independent component analysis (ICA) and multi-subject archetypal analysis (MSAA). 3) For the decomposition methods, we used three different transfer learning approaches (A1-A3) where information from the D1 feature extraction was transferred to extract similar features on the test dataset D2. 4) The prediction models were then trained either for diagnosis classification (using support vector machines (SVM)) and symptom regression (using Gaussian process regression (GPR)). 5) Finally, we validated the models on the data from the independent test site.

2.1 Participants and MRI acquisition

We used data from two publicly available datasets: i) 812 participants came from the DecNef Project Brain Data Repository (https://bicr-resource.atr.jp/srpbsopen/)[39], ii) 140 participants came from the Center of Biomedical Research Excellence (COBRE) dataset [40]. We split this into two datasets, the discovery dataset, D1 containing data from 8 sites (COBRE + 7 DecNef sites) which we use to compare and train different ML models and the independent test dataset, D2 containing data from two independent sites (both from the DecNef database) to assess the generalizability. With generalizability, we here refer to high and significant predictive performance of the biomarker on both the discovery and test datasets. With the aim to build a ML model that generalizes across datasets, we choose to keep data from as many subjects and sites as possible, which means that we have an unbalanced dataset with more healthy controls (HC) than patients with Schizophrenia (SZ). We made the splits between D1 and D2 such that approx. 70% of the data was used for training. For the prediction of the PANSS scores, we created a sub dataset, which included only data from SZ patients that had a PANSS score available, these datasets are referred to as D1a and D2a. Demographics are specified in Table 1. For all included subjects, we had a structural T1 weighted image, and 5-10 min resting state fMRI data with eyes open. More detailed information about the number of included participants and MRI acquisition parameters for each site can be found in Supplementary Table 10. This study was approved by the Institutional Ethical Review Board at the Technical University of Denmark, department for applied Mathematics and Computer Science (COMP-IRB-2022-03).

	Diagnosis classification				PANSS prediction	
	Discovery data (D1)		Test data (D2)		D1a	D2a
	HC	SZ	HC	SZ	SZ	SZ
n participant	486	143	260	63	136	44
n sites	8	3	2	2	3	1
Sex (♂,♀)	256/230	100/43	179/81	35/28	99/37	20/24
Age (μ,σ)	40 ±16	36 ±12	34 ±12	42 ±10	36 ±12	42 ±10
PANSS$_{total}$					62 ±17	57 ±18

Table 1) Participant demographics. Number of patients ($n_{participant}$) and sites (n_{site}), sex, age and PANSS total (measure of general symptom severity). The machine learning models were trained on dataset D1 and D1a and finally we tested the generalizability of the model using data from an independent test dataset, D2, including data from two independent sites. The suffix "a" indicates a subset of the datasets than only includes patients with an available PANSS score thus PANSS total is only listed for D1a and D2a.

2.2 Preprocessing

We aimed to keep our preprocessing pipeline as simple and generalizable as possible. We converted the raw datafiles into BIDS format and used fMRIprep v. 20.2.6 for preprocessing [22]. We used

fMRIpreps standard settings for slice timing correction, realignment, between modality registration, segmentation, and spatial normalization to standard space. Additionally, we used high pass filtering with a cut-off frequency of 1/128 Hz and 6mm FWHM isotropic Gaussian smoothing. For scans where B_0 field maps were available we estimated and applied a voxel displacement map based on the effective echo spacing and phase-encoding direction. "Fieldmap-less" distortion correction by matching the anatomical features to the T1-weigthed scan, was applied for scans where no B_0 field map was available [22]. We regressed out the mean signal of nuisance compartments (global, white matter and cerebrospinal fluid), 24 motion parameters [41], and scrubbed volumes where the framewise displacement exceeded 1mm. We excluded participants where more than 30% of volumes were scrubbed (nine participants). Please note that the excluded participants are not part of the number and demographics in section 2.1.

2.3 Feature extraction methods

We used three feature extraction methods; i) parcellation based connectivity analysis, ii) independent component analysis (ICA), and iii) the novel multi-subject archetypal analysis (MSAA).

Parcellation based connectivity analysis

Parcellation based connectivity analysis is the most used feature extraction method for neuroimaging data that is used for subsequent ML analysis. The parcels can be defined either as a sphere around a center coordinate or using different brain atlases. In this paper, we will also refer to parcels as regions of interests (ROI). We used the functionally defined 300-ROI set that was recently presented by Seitzman et al [42], which is an extension to the 264 ROI atlas from Power et al [43], where rsfMRI data was used to get an improved representation of ROIs in the subcortex and cerebellum [42]. The 300 ROI set can be downloaded on the Greene lab website. We excluded 25 ROIs that were mostly outside our group mask (threshold: <5 voxels within ROI), these were mainly located in the cerebellum (16/25) (which was not in the field of view for our scans) and in orbitofrontal regions plagued by signal dropout due to field inhomogeneity. We then used the MNI coordinates and sphere radius (5 or 4mm depending on the ROI) from the remaining ROIs and assigned the network labels of each ROI to the 7-network parcellation from Yeo et al [44], since this is the RSN parcellation that we have used for the decomposition methods. The seven RSN include: Visual, Somatomotor, dorsal attention (dATT), ventral attention (vATT), Limbic, Frontal-parietal (FPN) and Default mode network (DMN). Functional connectivities (FC) were calculated using Pearson's correlation coefficient with subsequent z scoring, which resulted in 37.675 FC (275 * (275-1))/2).

Group independent component analysis (ICA)

Group ICA is the most frequently used unsupervised ML method to extract brain networks from fMRI data, in particular, for the extraction of RSN from rsfMRI [29, 30]. This decomposition method aims to

identify a low rank representation of the data, such that the spatial sources are maximally independent. This means that ICA components are constructed such that the connectivity between components is minimized. We applied group ICA through the GroupICATv4.0a GIFT toolbox[45] using the Infomax algorithm, and the number of components (NOC) were selected using the minimum description length [46], which resulted in 23 components (median). We used a group mask that preserved voxels included in 95% of all individual brain masks (output from fMRIprep). Subject specific components were obtained using dual regression [47], as this allowed extraction of component expressions for independent datasets in a straightforward and consistent manner without data reduction procedures. For further analysis, components were z-scored. For visualization purposes we thresholded the activation at $|z|>1$.

Multi-subject archetypal analysis (MSAA)

MSAA is a novel decomposition method that aims to find characteristic archetypes in the data, which are latent factors that represent extremal points in the data [32]. For fMRI data, these archetypes are characteristic time-series, and a corresponding set of subject (denoted by i) specific spatial maps S_i [25]. Each map reflects the fractional contribution of each voxel to this archetype. Whereas ICA represents the fMRI data by a linear mixture of maximally independent spatial maps, MSAA determines the components through iterative optimization of; i) a seed region matrix, C (which is constrained to be identical for all subjects), and ii) a set of subject specific spatial maps (S_i) corresponding to each archetype. The archetypes for each subject are given as the weighted average of the voxels specified in the seed region matrix, such that $A_i = X_i C$, where X_i is the subject specific data and A_i includes all archetypes, defining distinct temporal profiles, for the i'th subject. The columns of both S_i and C are constrained to be nonnegative and to sum to one, which means that for each voxel, the time series is reconstructed by a convex combination as defined in S_i of the archetypes. The resulting spatial maps can therefore be interpreted as the fractional contribution of all voxels to the archetypal time series as specified in A_i. MSAA allows heteroscedastic noise modelling over voxels and subjects, such that the linear model per subject can be formulated as

$$X_i = X_i C S_i + E_i$$

Where the noise (E_i with columns $\varepsilon_{i,v}$) is assumed to be independently distributed with a Gaussian distribution such that $\varepsilon_{i,v} \sim N\left(0, I_T \sigma_{i,v}^2\right)$. Here $\sigma_{i,v}^2$ is the voxel (v) and subject specific noise variance and I_T a $T \times T$ identity matrix, where T is the number of timepoints. This leads to the following likelihood (L) function

$$L = \prod_i^B \prod_v^V \frac{1}{\left(2\pi\sigma_{i,v}^2\right)^{T/2}} \exp\left(\frac{-\|X_{iv} - X_i C S_{i,v}\|^2}{2\sigma_{i,v}^2}\right).$$

Similarly, as for other decomposition methods such as ICA, determining $\mathbf{C}$, $\mathbf{S}_i$, and $\sigma_{i,v}$ jointly leads to a nonconvex optimization problem (Mørup & Hansen, 2012), and a solution can be found by alternating optimization using projected gradient descent. For more details on MSAA and its relation to ICA, we refer to our previous publications [31, 33].

MSAA was implemented using the `MultiSubjectAA` code that is available via GitHub[48], and run using `Matlab` v. 2020b. We used the same group mask and number of components as for ICA (NOC = 23). We specified that the optimization should halt after either 1000 iterations or when the relative decrease in the cost function (L) was less than 10^{-6}, which was the same convergence criteria as used in our previous work [21, 25]. Since MSAA is a nonconvex optimization problem, there is a risk that the optimization identifies a local rather than the global minimum, and it is therefore suggested to repeat the analysis with several random initializations. In previous work, we specifically investigated the stability based on the number of initial repetitions and found that 10 repetitions increased the stability such that the same network were found for each run [33]. Here we therefore again repeated the analysis 10 times and chose the solution with the lowest final cost function (i.e., highest likelihood (L)).

Assigning components to RSN:

We assigned a RSN label to each component from ICA and MSAA by using the 7-RSN parcellation presented in Yeo et al. [44] . We used the same approach as in previous studies[49-51], such that a decomposition component was assigned to one of the 7 RSNs if its absolute mean spatial correlation (over participants) was > 0.2. In this way each component of the decomposition methods represents a sub-part of one of the Yeo networks and thus they mostly include connectivity information from within a single RSN, and less information about between RSN connectivity compared to the parcellation based analysis. E.g. the first two components for both ICA and MSAA both represent different parts of the visual cortex, as illustrated in Figure 3 and Supplementary Figure 1. The exact correlation values can be found in Supplementary Table 1. Decomposition components that had a correlation < 0.2% were regarded noise and hence not included in further analyses. Visual inspection of the discarded networks confirmed that these networks mainly included activation in non-neuronal tissue (e.g., in ventricles) or cerebellum which was outside the field of view for several sites, and therefore not included in this study. For both ICA and MSAA, 14 RSNs were found. We matched and compared the similarity between the RSNs extracted by each method, by using the absolute mean spatial correlation (over participants). The components were assigned the same name across methods if their spatial correlation exceeded 50% (which was the case for 11/14 networks).

2.4 Transfer learning

Since decomposition methods determine the brain networks which best explain the data, there is no guarantee that the same networks are found across datasets. However, this is essential to assess the generalizability of a prediction model on a new independent dataset. This is particularly challenging if the two datasets are very different either due to differences in the populations (e.g., different clinical (sub) populations, demographics etc.) or due to measurement differences (e.g., scanner type, protocol, experimental procedures etc.). In these cases, brain networks that were well-suited in the first dataset, might fail to identify features that are important in the other dataset, which in turn will hamper the generalizability of the prediction model. We therefore investigated three different ways to translate RSNs found in the discovery dataset (D1) to the independent dataset (D2). As in previous publications, we refer to these as "transfer learning approaches" to indicate that information from one decomposition is transferred to the next [52]. For all approaches, we used the same number of components as for the D1 dataset.

In **approach 1 (A1)** we simply rerun the decomposition analysis on the new dataset D2. The only "transfer learning" information is that we used the same number of components and algorithm settings as for D1.

In **approach 2 (A2)** we rerun the decomposition analysis on the merged dataset (D1+ D2). In this way, the decomposition has access to data from all participants, and if D1 >= D2, this is likely to influence the decomposition such that the components are more similar to the initial decomposition on D1. In practice this is the approach taken in most cross-validated ML prediction studies that employ decomposition for feature extraction; however, it may not be desirable since merging the datasets mean that the features of the two datasets are then no longer independent. It is important to emphasize that the decomposition method is not informed about the prediction label, and thus does not bias the prediction performance as such. However, it still violates the assumption of independence between datasets, which is of particular importance in our study as our D2 set only contains data from independent sites.

In approach 3 (A3) we directly use the output from the D1 decomposition to make the D2 decomposition. For ICA this is done by using the same dual regression procedure that was used to create subject specific spatial maps for D1. I.e., for each participant in D2, we again use dual regression with the D1 ICA decomposition map **S**. For MSAA, this was done by keeping the common seed generator matrix **C** from the D1 decomposition fixed for the D2 decomposition, and then allow the algorithm to make a few iterations until convergence, to optimize the subject specific spatial maps (**S**) and heteroscedastic noise estimations ($\sigma_{i,v}$) according to the new dataset.

To our knowledge, this is the first time that this transfer learning was applied to a MSAA decomposition. In principle A3 is the most straight forward option and ensures a direct matching between the components of the two datasets. However, if the datasets are different, components obtained with this approach will not adequately capture the information in the new dataset, which can lead to poor generalizability performance, simply because the extracted features are poorly defined.

A key challenge for A1 and A2 is that there is no guarantee that the same brain networks will be found (particularly if the datasets are different as described in the top section) and in these cases, it is difficult to match networks between the two datasets. We used a Procrustes alignment to match components between the datasets as done in earlier studies [33, 52], and confirmed all matches with visual inspection. In Supplementary table 2, we have listed the matching and correlations for each network, and visually shown all components that had a correlation below < 70% in Supplementary figure 2.

2.5 Diagnosis classification

For the diagnosis classification we use soft margin support vector classification (SVC), which has been applied in more than 50% of earlier neuroimaging classification studies in psychiatry [53, 54]. The SVC identifies a separating hyperplane that maximizes the margin (difference between the groups), this hyperplane is only defined by the support vectors which are datapoints from participants that are on the margin. The soft margin SVC allows misclassification to lower the risk of overfitting by introducing slack variables for each misclassified participant. We implemented SVC using the libSVM implementation in scikit-learn v. 1.0.2 in Python v. 3.9.10 [55, 56], using a linear kernel with C = 1 and balanced class weights. To ease computational costs, we precomputed all kernel evaluations. As performance measure we chose the area under the ROC curve (AUC), which is reasonably robust to label imbalance as exhibited in the current dataset. We trained the model on dataset D1 using 10-fold cross validation (CV) and estimated the classification certainty (i.e., how often the same label was predicted) when repeating the procedure 100 times with different data splits. For the ROI analysis and single RSN decomposition classification, the final diagnosis label was assigned based on the mean of the 100 repetitions (SZ if mean SVM prediction was > 0.5), for the decomposition ensemble classifier (more details below) we used the certainty estimate as weights in the soft voting. Finally, we tested the generalizability by applying the model on the dataset D2, which consisted of data from two independent sites.

For the ROI-based classification our main analysis used the whole connectivity matrix as the input feature, which measures how well the model classifies when using connectivity between all ROI pairs, as done in most previous studies.

Furthermore, we performed the following post-hoc analysis to investigate the importance of different parts of the connectivity matrix. First, we ran the classification using only "within RSN" connectivity,

i.e., connectivity between ROIs that all are within the same RSN. Secondly, we ran the classification only using "between RSN", i.e., keeping connectivity values between ROIs from different RSNs (e.g., between ROIs from DMN and FPN). And finally, we repeated the three analyses above (either keeping all, within or between RSN connectivity) for each individual RSN. In this way, we can evaluate the importance of each individual RSN. We preferred this approach over interpreting the classification weights, because interpretation of single weights are susceptible to misinterpretation[27, 57].

For the decomposition analysis (ICA and MSAA) we performed the following predictions:

- Classification on each individual RSN where we used the spatial map of each RSN as the input for the SVM classifier (14 separate classifications for both ICA and MSAA)
- Classification using all RSNs, using an ensemble decision across RSNs with a soft voting decision scheme. Here we used the certainty estimate of the 100 repetitions to weight each RSN, i.e., RSNs with a higher certainty

Classification performance on the validation dataset D2 was estimated on each of the three transfer learning approaches described in section 2.4, and for the ensemble decision the certainty weights from the training dataset were used.

Statistical significance was assessed using random permutation testing to obtain an empirical null distribution of the performance measure (AUC)[8, 58] . This was done by creating 1000 random permutations of the diagnosis label for each classification procedure (including all CV steps). For the classifications on individual RSNs from the decomposition methods, we corrected for multiple comparisons using maximum permutation statistics, i.e., we created an empirical null distribution by considering only the most significant effect over the entire set. This controls the family-wise error over the set.

2.6 Prediction of PANSS scores

Predictive assessment of the PANSS scores were performed using a Gaussian process regression (GPR) model utilizing a radial basis function kernel with automatic estimation of length scale and variance parameters via maximum likelihood estimation. GPR is a nonparametric, Bayesian approach to regression-based modelling, where the model infers a probability distribution over all possible values rather than attempting to predict an exact output. We chose to use GPR instead of linear support vector regression (SVR) (which previous studies have used [35-37]) since the PANSS scale (both total and subitems) is an ordinal scale, and because the GPR has some analytical advantages compared to SVR. Firstly, the GPR model allows hyperparameter estimation directly via the training likelihood, such that explicit hyperparameter tuning via cross validation is not needed. Secondly, instead of assigning an exact estimate of the PANSS score, it returns a probability distribution. In this way, the GPR returns a predicted PANSS score (mean) and a certainty estimate (standard deviation), which is a measure of

how confident the model is about its predictions. This can be extremely helpful when having to choose between different models, since it allows weighting more certain models higher.

To measure the prediction performance, we calculated the correlation between the predicted and observed PANSS scores. Since the PANSS scale is a summation of categorical subitems (and thus not a continuous measure), we used the Spearman's rank coefficient of correlations (Rho), which is a nonparametric measure of correlation utilizing ranks [59], and has less sensitivity to outliers (patients with very high/low PANSS scores). Furthermore, we also calculated the Pearson's correlation coefficient I, since this has been used in most of the earlier PANSS prediction studies[35-37], and therefore enables us to compare the performance to other studies.

For the parcellation based prediction, we performed one main analysis which included features from the whole connectivity matrix. For decomposition methods, we performed two analysis steps i) PANSS prediction on each individual RSN (as for the diagnosis classification) and ii) ensemble decision across all RSNs by comprising the probability functions for all individual RSNs as illustrated in Figure 2.

Our main goal was to predict the total PANSS score, which measure the overall symptom severity, the results of these findings will be discussed in section 3.4. Then we repeated the prediction using the three sub-subscales PANSS positive, negative, and generalized (discussed in section 3.5).

We implemented GPR using scikit-learn v. 1.0.2 in python v. 3.9.10, using a radial basis function kernel. On the discovery dataset, we initially made a short model comparison for the PANSS total prediction, where we tested different ML methods (both linear regressions, SVR and GPR) and kernels (linear, radial basis function and Matérn with smoothness parameters ν = 1.5) using dataset D1 and parcellation features. We found that most models performed similarly, results can be found in Supplementary Table 4. We continued to use GPR for the remaining analysis due to advantages described above.

As for the diagnosis classifications, the PANSS prediction models are trained on a multi-site dataset D1a and the generalizability is tested on data from an independent site, D2a. For the decomposition RSNs, we only tested the generalizability of RSNs that predicted significantly on dataset D1a. Overall the split between the datasets is the same as for the classifications; however, since we here only include patients with SZ, the datasets are smaller. The demographics are listed in Table 1.

For *statistical inference* we used the same permutation approach (with 1000 random permutations) as described in section 2.5. We accounted for multiple comparisons using maximum permutation statistics, i.e., only considering the most significant effect over the set of RSNs for each decomposition method.

For each patient:

Probability function for each RSN
μ = predicted PANSS
σ = certainty estimate

Posterior distribution (MOG)
Gaussian approximation

Ensamble prediction for each patient :
μ and σ of posterior Gaussian approximation
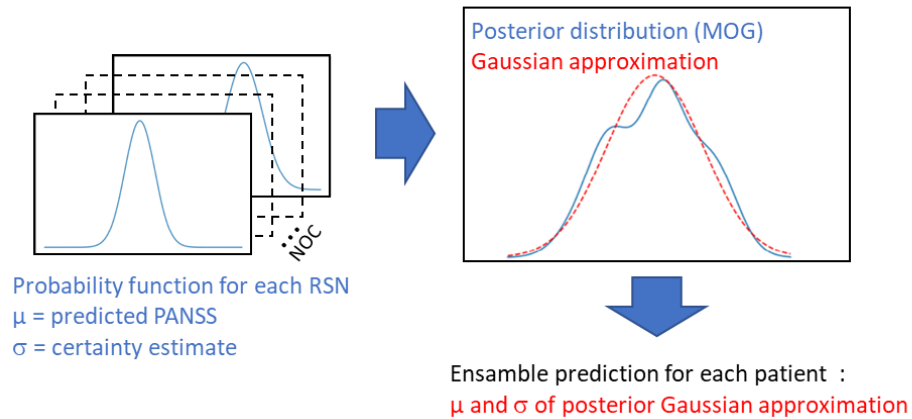
Figure 2: Ensemble prediction of GPR for decomposition methods. The ensemble prediction model for the PANSS prediction is built using the GPR outputs for all the individual RSNs. For each patient and RSN, the GPR outputs a probability density function which gives the predicted PANSS score (mean, μ) and certainty estimate (standard deviation, σ). To make an ensemble decision, we summed and normalized the probability functions of each RSN forming a Mixture of Gaussians (MOG) posterior distribution, and then fitted a Gaussian approximation to the resulting distribution to quantify the ensemble estimate and the associated uncertainty. The ensemble prediction for that patient is then given by the mean (predicted PANSS) and standard deviation (certainty estimate) for the Gaussian approximation.

Results

The results of our study are presented in the following five sections; 3.1: RSN extraction using decomposition methods, 3.2: classification of diagnosis, 3.3: transfer learning across datasets (for decomposition method), 3.4: prediction of symptom severity (PANSS total score), and 3.5: prediction of the PANSS subscales (positive, negative, and generalized).

3.1 RSN extraction using decomposition methods

ICA and MSAA both found 14 RSNs according to the 7-network parcellation presented in Yeo et al. [44]. The majority (11/14) of these networks were very similar between the two methods, and overall, we found that the MSAA RSNs were more distinctively expressed, as seen In Figure 3. This is particularly clear for the vATT network, which for MSAA includes strong bilateral expression in the frontal gyrus (inferior, medial, and superior), insula, superior temporal gyrus and inferior parietal lobule, whereas the vATT network for ICA included the same regions but to a lower extend. A more detailed visualization with several views for each RSN can be found in Supplementary Figure 1.
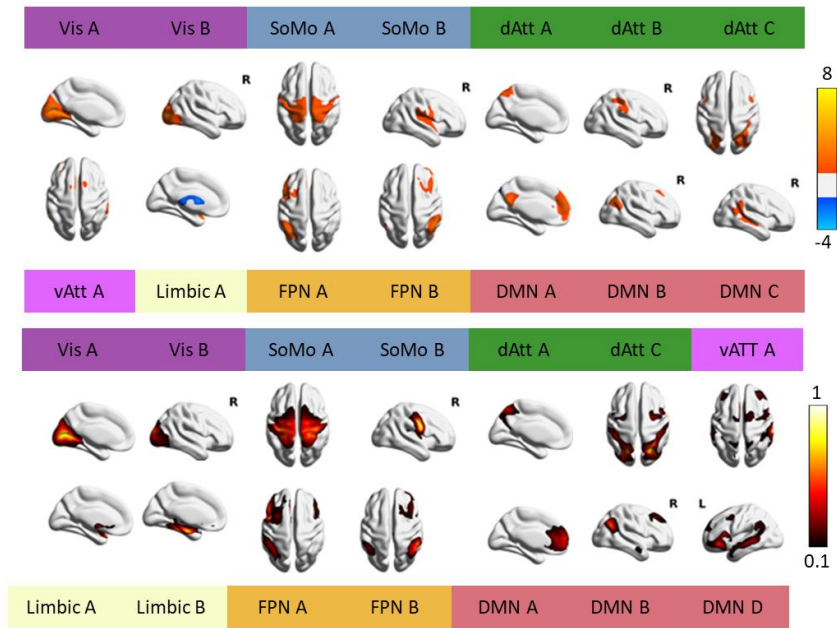
Figure 3: Visualization of resting state networks (RSN) from ICA (top) and MSAA (bottom). Decomposition components were categorized as RSN if their mean (over participants) correlation was > 0.2 to the RSN presented in the 7-network parcellation in Yeo et al [44] . For both ICA and MSAA, 14 RSN were found and networks where z-scored for further analysis. For visualization the ICA RSNs were cut off at $|Z| > 1$, and for MSAA networks include voxels with >10% fractional contribution. Vis: Visual, SoMo: Somatomotor, dATT: dorsal attention, vATT: ventral attention (also referred to as salience network), FPN: frontoparietal network, DMN: default mode network.

Transfer learning between datasets

Using the three transfer learning approaches (A1-A3 as described in section 2.4), we found that the RSN were stable with a mean (over participants and RSNs) spatial correlation of 96% for ICA and 98% for MSAA when using transfer learning approach A3. The mean stability for A1 and A2 were generally ~10% higher for MSAA (87% for both A1 and A2) than ICA (A1 = 73% and A2 = 79%). The mean stability for individual RSNs can be found in Supplementary Table 2. We performed a visual comparison of all networks that had a stability below 70% (Supplementary Figure 2), and found that even for these, the majority of larger brain regions included in each network were the same across the transfer learning approaches. High spatial correlations were also obtained for transfer learning approach A1 and A2, which showed that the extracted RSNs were also here relatively stable across datasets, but that signal strength and precise locations varied somewhat between datasets.

3.2 Classification of diagnosis label

For both the parcellation and decomposition analyses (ensemble classifiers, which included a soft voting scheme of all RSNs), we found a high classification performance with an AUC > 0.89 (p<0.001) on the discovery dataset D1, which generalized well to the independent validation set D2 with an AUC > 0.81 (p< 0.001). Overall classification performances were similar for all three methods, but the parcellation based features had the highest AUC, particularly on the validation dataset D2 as shown

in table 2 and Figure 4. For the decomposition methods, we found that transfer learning approach A3 yielded the best performance on the independent test dataset ($AUC_{ICA, D2} = 0.86$ and $AUC_{MSAA, D2} = 0.84$). This shows that this transfer learning approach was superior both regarding the spatial stability of the components (as described in section 3.1) and classification performance.

AUC	Discovery data set (D1)	Independent test set (D2)		
Parcellation connectivity	0.91	0.91		
Decomposition:		A1	A2	A3
ICA ensemble	0.90	0.84	0.81	0.86
MSAA ensemble	0.89	0.81	0.83	0.84

Table 2: Diagnosis classification performance of parcellation and ensemble decomposition features. ROC AUC for the parcellation based (ROI) analysis and ensemble classifiers (soft voting) for the decomposition methods. AUC listed for the discovery dataset D1 (using 10-fold CV) and on the validation test set D2 (data from two independent sites). For the decomposition methods, the performance on D2 is listed using all three transfer learning approaches, A1, A2 and A3. Performance of the individual RSNs can be found in supplementary table 3.
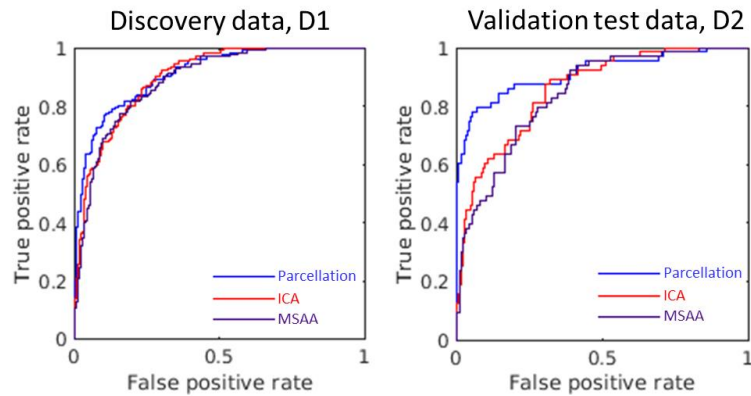
Figure 4: ROC curve for diagnosis classification. ROC curve for the D1 (left) and D2 (right) datasets, using features from parcellation based connectivity analysis (blue), ICA (red) and MSAA (purple). For the D2 data, only the ROC curve for transfer learning approach A3 is shown. Performance of the other transfer learning approaches can be found in Table 3 and Supplementary Table 1.

Post-hoc analysis for parcellation based connectivity analysis

We performed a post-hoc analysis to determine the separate importance of the different RSNs as well as the contributions of within and between RSN connectivity values. As shown in Figure 5, we found that the highest performances were obtained when using the whole connectome or the between RSN connectivity (AUC > 0.80 for all RSNs), while classifications that only used within RSN connectivity values obtained substantially lower performance (AUC: 0.56-0.86). Furthermore, the highest classification was obtained when using information from all RSNs ("ROI" columns to the left in Panel D and Panel E for Figure 5). For the individual RSNs, the highest classification performance was obtained with ROIs from the limbic RSN which yielded high classification performance (between RSN > 0.90, within RSN > 0.79) for both the discovery and validation test set. The other six RSNs also

obtained high and significant performance for the between RNS connectivity (> 0.80 on both datasets) while the within RSN connectivity values were somewhat lower, yet significant for most networks (0.56-0.82). A similar pattern (between RSN connectivity > within RSN connectivity and not much difference between RSNs) was also found when looking at the individual RSN contributions on the weight vectors (Supplementary Figure 3), which were transformed to forward inference for easier interpretation using the approach suggested by Haufe et al [26]. Finally, to see the directionality of these connectivity changes, we have shown the difference map between patients with SZ and healthy controls in Supplementary Figure 4. Here we found that patients with SZ have hypoconnectivity within the RSNs, whereas the between RSN connectivity is more mixed, including both hyper and hypoconnectivity.
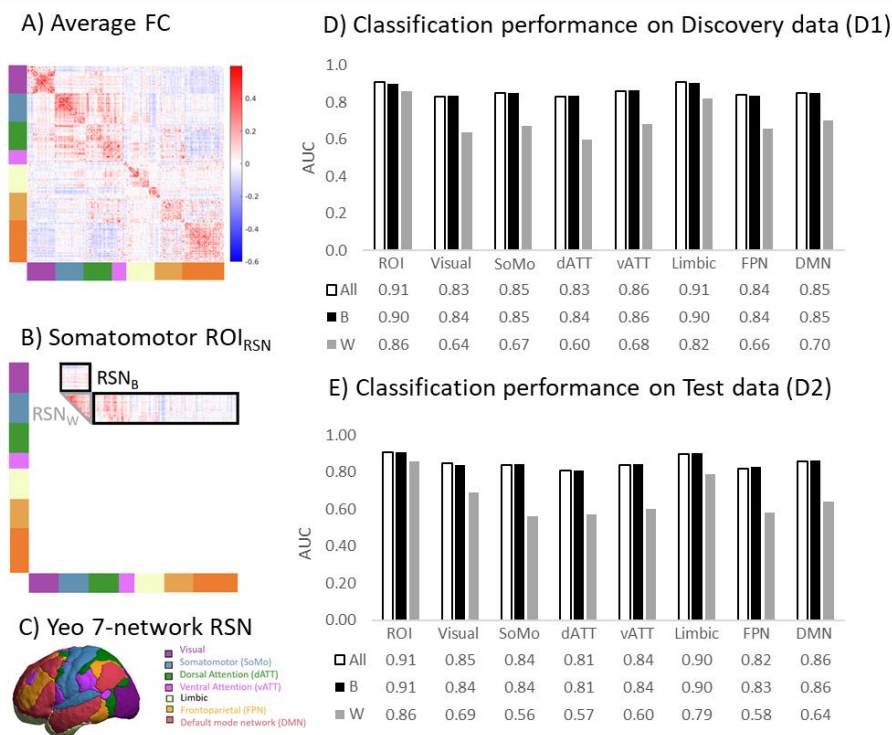
Figure 5: Post-hoc parcellation based analysis Panel A) visualization of the average (across participants) functional connectivity (FC) matrix for all ROIs, sorted according to the Yeo 7-network parcellation (Panel C). Panel B) visualization of Somatomotor RSN, where RSN_w refers within RSN connectivity for that RSN (grey) and RSN_B refers to the between network connectivity (black). RSN_{All} refers to all connectivity for the given network (white). The classification performance for each RSN is shown for the discovery (panel D) and independent test (panel E) dataset. Supplementary Figure 5 shown the Yeo 7-network parcellation in more details and views.

Classification on individual RSN for decomposition methods

When focusing on the classification of the individual RSNs, we see a similar trend as for the parcellation based post-hoc analysis. All networks resulted in significant classifications on the discovery dataset D1 ($AUC_{ICA,D1}$ = [0.65-0.81], $AUC_{MSAA,D1}$ = [0.67-0.81]), and most were also significant on the validation dataset D2 ($AUC_{ICA,D2}$ = [0.55-0.79], $AUC_{MSAA,D2}$ = [0.53-0.79]) when using transfer learning approach A3. The performances of all RSNs, datasets and transfer learning approaches are listed in

Supplementary Table 3. RSN within the somatomotor, visual, and dorsal attention network were consistently among the top three highest classifying networks for both ICA and MSAA on the discovery dataset. Furthermore, they obtained significant (yet not the best prediction performances compared to other RSNs) when tested on the independent test dataset.

3.4 Prediction of the PANSS total score

In this section we show the results from the total PANSS score ($PANSS_{total}$) which is a measure of the symptom severity. Figure 6 shows the performance for the parcellation based analysis (of all ROIs) and ensemble models for the decomposition methods. We found that the $PANSS_{total}$ prediction performance on the discovery dataset D1 was significant with moderate correlation (Rho > 0.3, r > 0.36, p < 0.001). However, when inspecting at the scatter plots in Figure 6, we see that even though the correlations between the observed and predicted PANSS score are significant, the predicted PANSS mostly reflect a trend around the mean (mean $PANSS_{D1a}$ = 62). This is also confirmed when looking at the predictions on the independent test dataset D2, where only the ICA features obtained significant classification, and even here the performance was modest (Rho = 0.27, r = 0.28, p = 0.01).
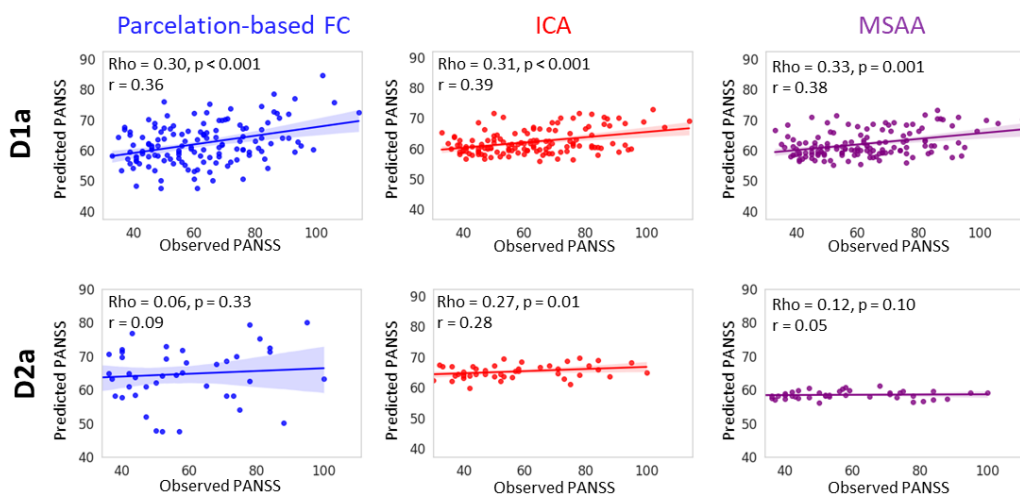
Figure 6: PANSS total prediction of ROI and ensemble models for decomposition. PANSS total prediction on the discovery dataset (D1a, top) and independent test set (D2a, bottom), using features from the parcellation analysis (blue, left), ICA (red, middle) and MSAA (purple, left). The primary prediction performance was Spearman's rank coefficient of correlation (Rho) and furthermore we also listed the Pearson's correlation coefficient (r), since this is mostly used in earlier studies. The scatter plots show the predicted PANSS (y-axis) as a function of the observed PANSS (x-axis), where the line shows the linear regression, and the shaded area indicates the standard error of the mean.

Predictions on individual RSNs

We did not find any RSN that obtained significant prediction on both datasets after correction for multiple comparisons. As shown in Supplementary Table 6, nine RSNs provided significant predictions on D1a but not on D2a. When comparing the performance of the two decomposition methods, both ICA and MSAA found that the highest test prediction performance for the ventral attention (vATT) RSN networks. For this RSN both methods found a low to moderate performance

(Rho$_{ICA}$ = 0.27 and Rho$_{MSAA}$ = 0.33, p< 0.05) on the discovery dataset, while the performance on the test dataset only showed a low (not significant after multiple comparisons correction) correlation (Rho$_{ICA}$ = 0.22 , Rho$_{MSAA}$ = 0.12, p$_{uncorrected}$ < 0.1). Visual inspection of the scatter plots of these predictions (Figure 7), show the same pattern as for the ensemble decision models, where all predictions are close to the mean PANSS score.
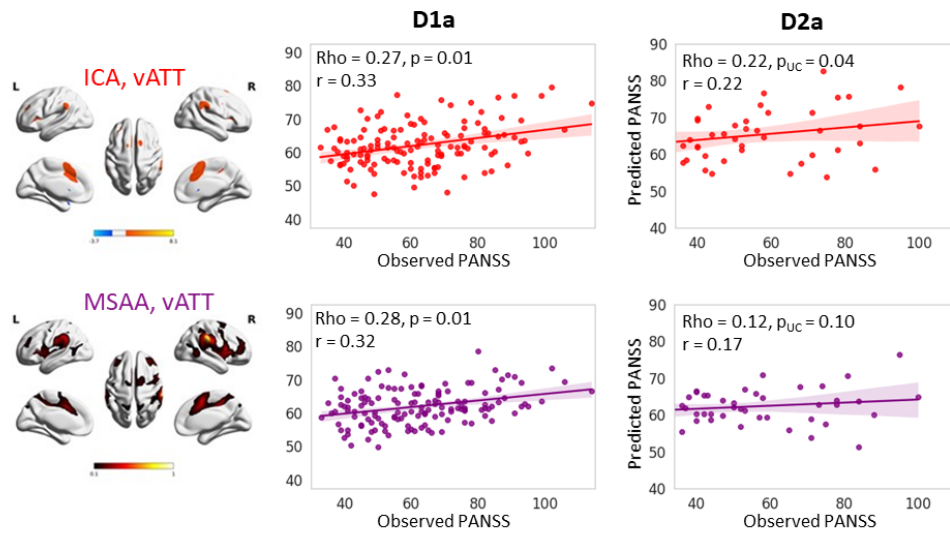
Figure 7: PANSS total prediction ventral attention (vATT) RSN for ICA and MSAA. Scatter plots of the vATT RSN, which obtained the highest prediction performance for both ICA (top, red) and MSAA (bottom, purple). Overall, the two vATT networks include similar brain regions, but the MSAA vATT network has stronger expressions in the bilateral frontal gyrus (inferior, medial, and superior), insula and superior temporal gyrus. The prediction performance on the discovery dataset D1a (middle) and test dataset D2a (right) are both listed with the Spearman's rank correlation coefficient (Rho) and Pearson's correlation coefficient (r). The shaded area of the linear regression line shows the standard error of the mean. Prediction performances of all RSNs that obtained significant prediction on D1a, can be found in Supplementary Table 6.

3.5 Prediction of PANSS subscales

In an attempt to disentangle the internal heterogeneity of schizophrenia, we also predicted the three PANSS subscales. For the parcellation-based features and ensemble decision of the decomposition methods, we found that the predictions were overall significant on the discovery dataset (apart from PANSS$_{positive}$ with ICA and MSAA features as listed in Table 3). However, as for the PANSS$_{total}$ predictions, the performance scores were low to moderate (Rho = [0.21-0.34]) and resembled a linear trend around the mean PANSS score. For the ICA model significant prediction was obtained for both the discovery and the test dataset for the generalized dimension (Table 3 and Figure 8), however with relatively low performances.

PANSS subscale	Positive				Negative				Generalized			
	D1a		D2a		D1a		D2a		D1a		D2a	
	rho	p	rho	p	rho	p	rho	p	rho	p	rho	p
Parcellation based	0.21	0.01	0.07	0.34	0.29	0.01	-0.03	0.53	0.25	0.01	0.07	0.48
ICA	0.18	0.15	0.31	0.01	0.28	0.01	0.07	0.27	0.20	0.03	0.24	0.04
MSAA	0.09	0.35	0.07	0.19	0.34	0.01	-0.10	0.93	0.27	0.02	0.16	0.28

Table 3: Prediction performance on three PANSS subscales using ROI and ensemble prediction models. Prediction performance for the three models that include information from all RSN (ROI and ensemble model for the decomposition methods ICA and MSAA). Significance assessed using random permutation statistics with multiple comparison correction over the three subscales. r = Pearson's correlation coefficient, D1a: SZ patients of discovery dataset, D2a: SZ patients of validation dataset.

As for the ensemble predictions, several **individual RSN** obtained significant classification on D1a, but the performances were low to moderate, and none of the predictions were significant when tested on the validation dataset (Supplementary table 8 and 9). In Panel B of Figure 8, we show the prediction result for the RSN (ventral attention (vATT A) from ICA) which obtained the highest performance on both datasets, however, the predictions provided by this RSN were not significant after multiple comparison correction.
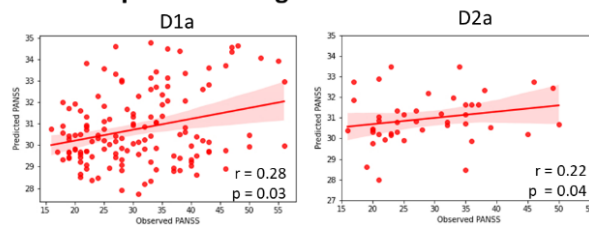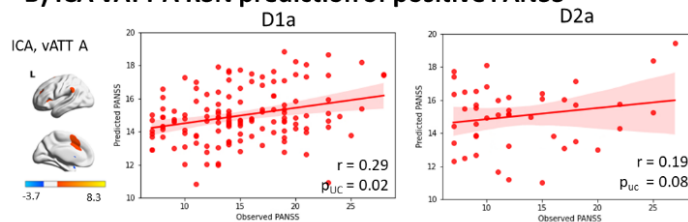
Figure 8 PANSS subscale predictions. Panel A) shows the ICA ensemble decision model which obtained significant prediction performance on the generalized dimension. Panel B) shows the PANSS positive prediction for the ventral attention (vATT A) RSN. This RSN from the ICA decomposition was the only RSN that obtained significant (uncorrected for multiple comparisons) classification on both the discovery (D1a) and test (D2a) dataset. No RSNs obtain significant classification of the other dimensions.

Discussion and future perspectives

In this section we will discuss our main findings as well as limitations and opportunities for future studies.

4.1 Decomposition methods for feature extraction across datasets

Data-driven feature extraction methods, such as ICA and MSAA, extract brain features (networks or regions) that are characteristic for the given dataset, which can both be an advantage as well as represent challenges. On the one hand, the features extracted with these methods represent strong trends for the specific datasets, and the networks can be more sensitive compared to parcellation based approaches, where the atlas does not necessarily fit the brain of the participant very well [60]. On the contrary, it also means that the extracted features can be overfitted to the specific dataset. This can lead to lower generalizability, and there is no guarantee that the same networks are found again when rerunning the analysis on another dataset[1]. To investigate these challenges, we compared three different transfer learning approaches which bridged networks from our discovery dataset (D1) and test dataset (D2). In summary these included, A1) rerun analysis on the new dataset (D2), A2) rerun analysis on the pooled dataset (D1 + D2) and A3) directly use the decomposition from D1 to construct the features of D2.

First of all, on the **discovery dataset** we found that both ICA and MSAA found 14 RSNs according to the 7-network parcellation presented in Yeo et al. [44]. Most of the networks were very similar between the two methods, and overall, the activation was more strongly expressed for MSAA networks as shown in Figure 3. From the investigation of **network stability across datasets**, we found that the highest spatial stability was obtained with transfer learning approach A3, where both decomposition methods extract similar RSNs that are stable (> 90%) across datasets. For both A1 and A2, we also found a high mean stability across networks, which were generally ~10% higher for MSAA compared to ICA. This showed that the RSNs were stable across datasets, but that signal strength and precise activations varied somewhat between sites. Furthermore, for A1 and A2 it was a challenge that there was no direct coupling of the RSNs between the datasets, and there was no guarantee that the same networks were found (as described in section 3.1 and supplementary Figure 2).

Overall, we found that both decomposition methods extracted stable and similar RSNs on the multi-site data, which were consistent with what has previously been presented in other single site studies. RSN extracted with MSAA were more distinctively expressed and obtained somewhat higher stability across datasets than ICA. For both ICA and MSAA, transfer learning approach A3 provided the most stable networks across datasets. Furthermore, since this method also enables a direct matching of

---

[1] This is even the case when repeating the analysis on the same dataset, due to the non-convexity of the optimization problems. This and the mitigation strategy are further elaborated in section 2.3.

components and a reduced computational complexity, we believe that this transfer learning approach is very promising.

4.2 Diagnosis classification

For our prediction analysis, we started with a diagnosis classification since this enables us to compare the utility of our multi-site prediction framework including the use of decomposition methods for feature extraction with previous literature (which have mostly focused on diagnosis classification [5]). We found that high and significant classification performance was obtained when using both the features from the parcellation based connectivity analysis and the decomposition method ($AUC_{D1} >$ 0.89) with similar performances as for earlier single-site classification studies [7, 54]. The best performances were obtained when using information from all RNS (whole connectivity matrix for parcellation based connectivity analysis and ensemble decision for decomposition methods), and these findings also generalized to the validation test dataset (D2). Overall, the performances were similar, but the classifications using the parcellation based features had the highest AUC, particularly on the test dataset D2 as shown in table 2 and Figure 4.

From the performances of individual RSNs, we still found significant classification with most networks, but we did find any "single best" RSN with substantially higher performance than the remaining networks. This was the case for both the classifications of the individual RSNs from the decomposition methods, and for the post-hoc analysis of the parcellation based connectivity analysis. Furthermore, our post-hoc analysis showed that *between* RSN connectivity features were of great importance, which could be a potential explanation for why the parcellation based connectivity features obtained higher performances that the ensemble decomposition methods (particularly on the test dataset D2), as the decomposition methods mainly included information about within RSN connectivity changes.

Taken together, these findings show that the best performances were obtained when using information from all RSNs, which is in accordance with earlier studies that have shown that patients with schizophrenia have affected connectivity changes in a large part of the connectome [14, 16].

However, as described in the introduction, the usefulness of diagnostic biomarkers in SZ have been challenged, since it could be argued that they have limited clinical utility [4, 25]. Instead, it has been suggested that it might be more fruitful to use fMRI to search for more biologically homogeneous subgroups [4, 61], which is why the second part of our prediction analysis was centered on predicting the PANSS scores of the patients.

4.2 PANSS predictions

In the PANSS predictions we used the same brain features as for the classification analysis, but we now used regression-based prediction to predict the total symptom severity ($PANSS_{total}$) and three PANSS subscales (positive, negative, and generalized), to address the internal heterogeneity of schizophrenia.

For both the PANSS$_{total}$ and subscale predictions, the performances were low to moderate, even for the discovery dataset, where the predictions resembled a positive trend around the mean PANSS score, and overall, the findings did not reproduce on the test dataset. For the decomposition method we found that the ensemble models had a better performance than the predictions of the individual RSNs, just as for the classification analysis. For the parcellation based connectivity analysis we opted against performing a post-hoc analysis since the predictions on the whole connectivity matrix were already limited and did not reproduce to the validation test set D2a.

For the individual RSN prediction of the PANSS$_{total}$ scores we found nine RSNs that predicted significantly on D1a but not on D2a, which highlights the importance of validating prediction models in an independent test dataset, to see if a prediction model, and thereby potential biomarker is reproducible to data from independent sites [23].

Furthermore, for both the PANSS$_{total}$ and PANSS subscale predictions, we found that it were not the RSNs that obtained highest performances on D1a, which obtained the best performances on the test dataset D2a. For example, for the PANSS$_{total}$ prediction, one of the ICA networks (DMN B) obtained the highest prediction performance on D1a (Rho = 0.34, =0.01) but a low performance on D2a (Rho = 0.12, $p_{uncorrected}$ = 0.12). On the contrary, the ventral attention network (vATT A) had a significant, but lower performance D1a (Rho = 0.27, p=0.05) but still showed a similar correlation on D2a (Rho = 0.22, $p_{uncorrected}$ = 0.04). This supports the "multiple comparison paradox" that was recently described in Marek et al. [62], where they found that correcting for multiple comparisons reduced the probability of successfully replicating univariate brain-wide association studies (BWAS). More specifically, they found that using a stringent statistical threshold (thereby selecting most strong BWAS effects), reduces the false positive rates, but increases the false negative rates and thereby lowers the statistical power. In underpowered studies, these strict statistical thresholds enforce detection of very large correlations, which are the most likely to be inflated due to sampling variability [62].

This poor generalizability of the predictions indicates that the study might have been underpowered [62] or that differences between sites were too large compared to the signal of interest (this is further discussed in section 4.4). Another explanation could be the internal consistency of the PANSS subscales itself, where items within a subscale have shown modest internal consistency [63], while scores tend to be correlated across subscales [64]. If the individual items of the PANSS scale had been available, it would have been interesting to see if those individual PANSS item predictions would be higher and more stable across datasets. Finally, in the light of our results, we also consider that our applied method (supervised machine learning) or even the imaging modality itself (connectivity in rsfMRI), might not be the right path forward to find robust biomarkers for symptom severity in

schizophrenia. In the following section we will therefore discuss the limitations of our work, and our suggestion for future studies.

4.4 Limitations and suggestions for future studies

There are several limitations of our work, which we would here like to discuss together with our recommendations for future studies.

First of all, even though we used a multi-site rsfMRI dataset, the number of schizophrenia patients was still not very high. For the classification analysis, we found that the sample size was sufficient to enable high and significant classification, which was reproducible across datasets. However, this was not found for PANSS prediction analysis. Regardless of the datatype, a more complex machine learning model requires more data to train a robust model. However, in our study we had much fewer participants for the PANNS regression analysis, since we here had to exclude data from healthy controls and patients that did not have a PANSS score available. This reduced the sample size to 136 for the discovery and 44 on the test dataset. Furthermore, since these patients came from five different sites, we only had 19-55 patients with schizophrenia from each site (specific numbers for each site listed in supplementary table 10). Multi-site data will typically introduce both measurement and sampling bias into the data, if the sample size if sufficient, this can be an advantage as the final model will then be more robust to between-site differences[6, 9, 65]. However, with the relatively low sample sizes from each site, these biases can also introduce too much variability to enable meaningful prediction, particularly if the site differences are correlated with the outcome measures[66, 67]. We therefore believe that a study with a larger schizophrenia sample size, and more patient from each site are needed in future studies to make firm conclusions on the ability of the applied methods for clinically meaningful and reliable PANSS predictions.

In future studies, we also suggest to investigate if multi-site variability reduction methods, with[67, 68] or without [69] travelling subject, can help to get better prediction by removing "measurement bias" from the data. Whereas some earlier studies have shown that between-site measurement bias can be implicitly handled by the machine learning model when sufficient data is included [65], multi-site variability reduction might be useful, and potentially even needed if it is not possible to obtain a large enough sample.

Another challenge of our study was the limited information about confounding factors for most of our participants, such as lifestyle specifications and treatment history. As for the between site biases described above, it can be an advantage not to remove confounding factors (either by removing observations or regressing factor out) but to let the model learn from the non-disease related heterogeneity to get a more generalizable outcome[9], which is in line with a general trend in machine

learning towards "more learning, less cleaning"[70]. However, since some of these factors have systematic differences between groups (such as lifestyle differences in smoking[71]), we would have preferred to have more information about confounding factors available to determine to what degree they influence our results. Similarly, for the PANSS prediction results, where we had a smaller dataset, it would have been interesting to investigate how factors such as disease duration, treatment history or clinical state of patients at the time of the evaluation would affect our results.

In consistency with most earlier studies, we used supervised machine learning to perform our predictions, but recently many studies also started to investigate the use of deep learning methods for neuroimaging predictions. In a recent review from Sadeghi et al, they summarize the increasing number of deep learning methods used for diagnosis classification in schizophrenia, and it will be interesting to see if these methods can obtain better prediction of PANSS scores or other symptom domains. Another opportunity would be to use unsupervised machine learning methods to search for subgroups with a more homogenous biology without the use of any "labelling" such as diagnosis or PANSS scores. Whereas earlier subtyping studies in schizophrenia have had some challenges from a methodological point of view [72], they have shown the important potential of using unsupervised subtyping in schizophrenia [73, 74]. Given the recent advancements in clustering methods that are specifically developed for high dimensional data such as rsfMRI[72, 75, 76], we believe that this is a promising avenue for future studies, with high potential clinical utility.

Finally, we also see a great potential for multi-modal studies, which combine different kinds of data, such as rsfMRI with structural imaging for their prediction analysis. Since all imaging modalities have different advantages and challenges, multi-modal methods carry a great potential to combine "the best of two worlds" if they can be implemented appropriately. For example, earlier studies have shown promising results when combining multimodal fusion of MRI data for PANSS prediction, where they obtained a high prediction performance (r > 0.7 on PANSS positive and negative) [77] .

Conclusion

In this work, we used one of the largest available multi-site rsfMRI datasets in schizophrenia and applied machine learning for both data-driven feature extraction (using the two decomposition methods ICA and MSAA) and prediction of the diagnosis labels and PANSS symptom scores. Comparing ICA and MSAA, we found that both methods extract similar RSNs which were stable across the two datasets. Using these RSNs and features from a parcellation based analysis, we demonstrated that classification models trained on multi-site fMRI data could significantly classify patients with schizophrenia from healthy controls with a high performance, and which reproduced on the independent test dataset. As in earlier studies we did not find any "single best" RSN that drove this

classification, but rather that changes within and between most RSNs were important for robust classification. When using the same features to predict the symptom severity and the three PANSS subscales, we did not find any clinically relevant predictions, since even the performances on the discovery dataset were low to moderate, and the models generally did not generalize to the independent test data.

We see our work as an important step towards building robust pipelines that combine multi-site rsfMRI data with machine learning method, and we hope that the data sharing initiatives will continue and expand, as we believe that even more multi-site data from patients, including information about confounding factors, are needed to make firm conclusion on the biomarker potential using these methods.

Acknowledgement:

Data and code availability.

The raw data is available through the DecNef and COBRE databases as described in section 2.1

We are happy to share resulting maps from our decompositions (**S** matrix for ICA and **C** for MSAA) given that access has been obtained to the databases as required by DecNef and COBRE. The Code for MSAA can be found on Github (https://github.com/JesperLH/Multisubject-Archetypal-Analysis).

References:

1. American Psychiatric, A., *Diagnostic and Statistical Manual of Mental Disorders, DSM-5*. 2013: American Psychiatric Association.

2. World Health, O., *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. 2004, World Health Organization: Geneva.

3. Kay, S.R., A. Fiszbein, and L.A. Opler, *The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia.* Schizophrenia Bulletin, 1987. **13**(2): p. 261-276.

4. Kraguljac, N.V., et al., *Neuroimaging Biomarkers in Schizophrenia.* Am J Psychiatry, 2021. **178**(6): p. 509-521.

5. Woo, C.-W., et al., *Building better biomarkers: brain models in translational neuroimaging.* Nature Neuroscience, 2017. **20**(3): p. 365-377.

6. Calhoun, V.D., G.D. Pearlson, and J. Sui, *Data-driven approaches to neuroimaging biomarkers for neurological and psychiatric disorders: emerging approaches and examples.* Curr Opin Neurol, 2021. **34**(4): p. 469-479.

7. de Filippis, R., et al., *Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review.* Neuropsychiatr Dis Treat, 2019. **15**: p. 1605-1627.

8. Arbabshirani, M.R., et al., *Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls.* Neuroimage, 2017. **145**(Pt B): p. 137-165.

9. Schnack, H.G., *Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases).* Schizophr Res, 2019. **214**: p. 34-42.

10. Orban, P., et al., *Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity.* Schizophrenia Research, 2018. **192**: p. 167-171.

11. Yoshihara, Y., et al., *Overlapping but Asymmetrical Relationships Between Schizophrenia and Autism Revealed by Brain Connectivity.* Schizophrenia Bulletin, 2020. **46**(5): p. 1210-1218.

12. Friston, K., et al., *The dysconnection hypothesis (2016).* Schizophrenia research, 2016. **176**(2-3): p. 83-94.

13. Li, S., et al., *Dysconnectivity of Multiple Brain Networks in Schizophrenia: A Meta-Analysis of Resting-State Functional Connectivity.* Frontiers in Psychiatry, 2019. **10**.

14. Pettersson-Yeo, W., et al., *Dysconnectivity in schizophrenia: Where are we now?* Neuroscience & Biobehavioral Reviews, 2011. **35**(5): p. 1110-1124.

15. Stephan, K.E., K.J. Friston, and C.D. Frith, *Dysconnection in Schizophrenia: From Abnormal Synaptic Plasticity to Failures of Self-monitoring.* Schizophrenia Bulletin, 2009. **35**(3): p. 509-527.

16. Dong, D., et al., *Dysfunction of Large-Scale Brain Networks in Schizophrenia: A Meta-analysis of Resting-State Functional Connectivity.* Schizophrenia Bulletin, 2018. **44**(1): p. 168-181.

17. Schnack, H.G. and R.S. Kahn, *Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters.* Front Psychiatry, 2016. **7**: p. 50.

18. Van Dijk, K.R., et al., *Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization.* J Neurophysiol, 2010. **103**(1): p. 297-321.

19. Noble, S., D. Scheinost, and R.T. Constable, *A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis.* NeuroImage, 2019. **203**: p. 116157.

20. Noble, S., D. Scheinost, and R.T. Constable, *A guide to the measurement and interpretation of fMRI test-retest reliability.* Current Opinion in Behavioral Sciences, 2021. **40**: p. 27-32.

21. Noble, S., et al., *Multisite reliability of MR-based functional connectivity.* NeuroImage, 2017. **146**: p. 959-970.

22. Esteban, O., et al., *fMRIPrep: a robust preprocessing pipeline for functional MRI.* Nature Methods, 2019. **16**(1): p. 111-116.

23. Scheinost, D., et al., *Ten simple rules for predictive modeling of individual differences in neuroimaging.* NeuroImage, 2019. **193**: p. 35-45.

24. FDA-NIH. *BEST (Biomarkers, EndpointS, and other Tools) Resource*. 2016.

25. Mechelli, A. and S. Vieira, *From models to tools: clinical translation of machine learning studies in psychosis.* npj Schizophrenia, 2020. **6**(1).

26. Haufe, S., et al., *On the interpretation of weight vectors of linear models in multivariate neuroimaging.* NeuroImage, 2014. **87**: p. 96-110.

27. Shen, X., et al., *Using connectome-based predictive modeling to predict individual behavior from brain connectivity.* Nature Protocols, 2017. **12**(3): p. 506-518.

28. Beckmann, C.F., et al., *Group comparison of resting-state FMRI data using multi-subject ICA and dual regression.* Neuroimage, 2009. **47**(Suppl 1): p. S148.

29. De Luca, M., et al., *fMRI resting state networks define distinct modes of long-distance interactions in the human brain.* Neuroimage, 2006. **29**(4): p. 1359-1367.

30. Smith, S.M., et al., *Resting-state fMRI in the Human Connectome Project.* NeuroImage, 2013. **80**: p. 144-168.

31. Hinrich, J.L., et al., *Archetypal analysis for modeling multisubject fMRI data.* IEEE journal of selected topics in signal processing, 2016. **10**(7): p. 1160-1171.

32. Mørup, M. and L.K. Hansen, *Archetypal analysis for machine learning and data mining.* Neurocomputing, 2012. **80**: p. 54-63.

33. Krohne, L.G., et al., *Classification of social anhedonia using temporal and spatial network features from a social cognition fMRI task.* Human Brain Mapping, 2019. **40**(17): p. 4965-4981.

34. Cheng, W., et al., *Voxel-based, brain-wide association study of aberrant functional connectivity in schizophrenia implicates thalamocortical circuitry.* npj Schizophrenia, 2015. **1**(1): p. 15016-15016.

35. Wang, D., et al., *Individual-specific functional connectivity markers track dimensional and categorical features of psychotic illness.* Molecular Psychiatry, 2020. **25**(9): p. 2119-2129.

36. Fan, Y.S., et al., *Tracking Positive and Negative Symptom Improvement in First-Episode Schizophrenia Treated with Risperidone Using Individual-Level Functional Connectivity.* Brain Connect, 2022. **12**(5): p. 454-464.

37. Fan, Y.S., et al., *Individual-specific functional connectome biomarkers predict schizophrenia positive symptoms during adolescent brain maturation.* Human Brain Mapping, 2021. **42**(5): p. 1475-1484.

38. Koch, S.P., et al., *Diagnostic classification of schizophrenia patients on the basis of regional reward-related fMRI signal patterns.* PLoS ONE, 2015. **10**(3).

39. Tanaka, S.C., et al., *A multi-site, multi-disorder resting-state magnetic resonance image database.* Scientific Data, 2021. **8**(1).

40. (COBRE), T.C.f.B.R.E.; Available from: http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.

41. Friston, K.J., et al., *Movement-related effects in fMRI time-series.* Magn Reson Med, 1996. **35**(3): p. 346-55.

42. Seitzman, B.A., et al., *A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum.* NeuroImage, 2020. **206**: p. 116290.

43. Power, J.D., et al., *Functional network organization of the human brain.* Neuron, 2011. **72**(4): p. 665-78.

44. Yeo, B.T., et al., *The organization of the human cerebral cortex estimated by intrinsic functional connectivity.* J Neurophysiol, 2011. **106**(3): p. 1125-65.

45. Rachakonda, S., et al., *Group ICA of fMRI toolbox (GIFT) manual.* Dostupnez [cit 2011-11-5], 2007.

46. Li, Y.O., T. Adalı, and V.D. Calhoun, *Estimating the number of independent components for functional magnetic resonance imaging data.* Human brain mapping, 2007. **28**(11): p. 1251-1266.

47. Smith, S.M., et al., *Correspondence of the brain's functional architecture during activation and rest.* Proceedings of the National Academy of Sciences, 2009. **106**(31): p. 13040-13045.

48. Hinrich, J. *Multi-subject Archetypal Analysis algortihm*. 2022 [cited 2022 2022]; Available from: https://github.com/JesperLH/Multisubject-Archetypal-Analysis.

49. Reineberg, A.E., et al., *Resting-state networks predict individual differences in common and specific aspects of executive function.* NeuroImage, 2015. **104**: p. 69-78.

50. Champagne, A.A., et al., *Multi-modal normalization of resting-state using local physiology reduces changes in functional connectivity patterns observed in mTBI patients.* NeuroImage: Clinical, 2020. **26**: p. 102204.

51. Van Timmeren, T., et al., *Connectivity networks in gambling disorder: a resting-state fMRI study.* International Gambling Studies, 2018. **18**(2): p. 242-258.

52. Cai, X.L., et al., *Generalizability of machine learning for classification of schizophrenia based on resting-state functional MRI data.* Human Brain Mapping, 2020. **41**(1): p. 172-184.

53. Sadeghi, D., et al., *An overview of artificial intelligence techniques for diagnosis of Schizophrenia based on magnetic resonance imaging modalities: Methods, challenges, and future works.* Computers in Biology and Medicine, 2022. **146**: p. 105554.

54. Steardo, L., Jr., et al., *Application of Support Vector Machine on fMRI Data as Biomarkers in Schizophrenia Diagnosis: A Systematic Review.* Front Psychiatry, 2020. **11**: p. 588.

55. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM Trans. Intell. Syst. Technol., 2011. **2**(3): p. Article 27.

56. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**(null): p. 2825–2830.

57. Haufe, S., et al., *On the interpretation of weight vectors of linear models in multivariate neuroimaging.* NeuroImage, 2014. **87**: p. 96-110.

58. Nichols, T.E. and A.P. Holmes, *Nonparametric permutation tests for functional neuroimaging: a primer with examples.* Human brain mapping, 2002. **15**(1): p. 1-25.

59. *Spearman Rank Correlation Coefficient*, in *The Concise Encyclopedia of Statistics*. 2008, Springer New York: New York, NY. p. 502-505.

60. Dadi, K., et al., *Benchmarking functional connectome-based predictive models for resting-state fMRI.* NeuroImage, 2019. **192**: p. 115-134.

61. Miranda, L., et al., *Systematic Review of Functional MRI Applications for Psychiatric Disease Subtyping.* Frontiers in Psychiatry, 2021. **12**.

62. Marek, S., et al., *Reproducible brain-wide association studies require thousands of individuals.* Nature, 2022. **603**(7902): p. 654-660.

63. Peralta, V. and M.J. Cuesta, *Psychometric properties of the Positive and Negative Syndrome Scale (PANSS) in schizophrenia.* Psychiatry Research, 1994. **53**(1): p. 31-40.

64. Van den Oord, E.J.C.G., et al., *Factor structure and external validity of the PANSS revisited.* Schizophrenia Research, 2006. **82**(2): p. 213-223.

65. Abraham, A., et al., *Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example.* NeuroImage, 2017. **147**: p. 736-745.

66. Yamashita, A., et al., *Generalizable brain network markers of major depressive disorder across multiple imaging sites.* PLOS Biology, 2020. **18**(12): p. e3000966.

67. Yamashita, A., et al., *Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias.* PLoS Biol, 2019. **17**(4): p. e3000042.

68. Yoshida, K., et al., *Prediction of clinical depression scores and detection of changes in whole-brain using resting-state functional MRI data with partial least squares regression.* PLOS ONE, 2017. **12**(7): p. e0179638.

69. Yu, M., et al., *Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data.* Human Brain Mapping, 2018. **39**(11): p. 4213-4227.

70. Cvetkov-Iliev, A., A. Allauzen, and G. Varoquaux, *Analytics on Non-Normalized Data Sources: More Learning, Rather Than More Cleaning.* IEEE Access, 2022. **10**: p. 42420-42431.

71.   Moran, L.V., et al., *Neural Responses to Smoking Cues in Schizophrenia.* Schizophr Bull, 2018. **44**(3): p. 525-534.

72.   Miranda, L., et al., *Systematic Review of Functional MRI Applications for Psychiatric Disease Subtyping.* Front Psychiatry, 2021. **12**: p. 665536.

73.   Brodersen, K.H., et al., *Dissecting psychiatric spectrum disorders by generative embedding.* NeuroImage: Clinical, 2014. **4**: p. 98-111.

74.   Yang, Z., et al., *Brain Network Informed Subject Community Detection In Early-Onset Schizophrenia.* Scientific Reports, 2015. **4**(1).

75.   Khosla, M., et al., *Machine learning in resting-state fMRI analysis.* Magn Reson Imaging, 2019. **64**: p. 101-121.

76.   Bouveyron, C. and C. Brunet-Saumard, *Model-based clustering of high-dimensional data: A review.* Computational Statistics & Data Analysis, 2014. **71**: p. 52-78.

77.   Meng, X., et al., *Predicting individualized clinical measures by a generalized prediction framework and multimodal fusion of MRI data.* NeuroImage, 2017. **145**: p. 218-229.

Supplementary material

In the following we have added additional figure and tables that were not included in the main manuscript.

Contents

Supplementary Figure 1: Detailed view of all RSNs

Supplementary Figure 1: Visualization of Resting state network (RSN) extracted with independent component analysis (ICA) and multi subject archetypal analysis (MSAA). For each, 14 RSN were found. Vis: Visual, SenMo: Somatomotor, dATT: dorsal attention, vATT: ventral attention (also referred to as salience network), Lim: limbic, FPN: fronto parietal network, DMN: default mode network, corresponding to the Yeo 7-network parcellation. Components were assigned to a network if their mean correlation was above 0.2. The RSNs were mapped on the cortical surface by using the BrainNet Viewer package (http://www.nitrc.org/projects/bnv).

Supplementary Table 1: RSN correlation to Yeo 7-network parcellation

ICA		MSAA	
RSN name	Correlation	RSN name	Correlation
Vis A	0.6	Vis A	0.6
Vis B	0.6	Vis B	0.5
SoMo A	0.5	SoMo A	0.6
SoMo B	0.3	SoMo B	0.4
dAtt A	0.2	dAtt C	0.5
dATT B	0.2	dAtt A	0.2
dATT C	0.4	vAtt A	0.6
vATT A	0.4	Lim B	0.3
Lim A	0.2	Lim A	0.3
FPN A	0.2*	FPN B	0.4
FPN B	0.3	FPN A	0.4
DMN A	0.5	DMN A	0.5
DMN B	0.3	DMN B	0.4
DMN C	0.4	DMN D	0.3
Mean	0.42	Mean	0.38

Supplementary table 1: Resting state network (RSN) correlation to Yeo. Mean absolute correlation (over participants) of decomposition components to the 7-network parcellation presented in Yeo et al [3]. Networks were categorized to a Yeo RSN if their correlation was >0.2. * For this RSN, there was a mean correlation > 0.2 for both FPN (0.22) and DMN (0.21). Here the assignment was based on the highest correlation, namely FPNS. For both independent component analysis (ICA) and multi subject archetypal analysis (MSAA), 14 RSN were found. Vis: Visual, SoMo: Somatomotor, dATT: dorsal attention, vATT: ventral attention (also referred to as salience network), Lim: limbic, FP: fronto parietal, DMN: default mode network.

Supplementary Table 2: Stability of RSN across transfer learning

		Independent component analysis (ICA)					Multi-subject archetypal analysis (MSAA)		
		D1	D2				D1	D2		
			A1	A2	A3			A1	A2	A3
Ind	RSN		Corr	Corr	Corr	Ind	RSN	Corr	Corr	Corr
1	Vis A		0.96	0.97	0.98	1	Vis A	0.97	0.98	0.99
2	Vis B		0.87	0.95	0.97	2	Vis B	0.85	0.89	0.97
3	SoMo A		**0.72**	0.86	0.96	3	SoMo A	0.97	0.98	0.99
4	SoMo B		0.87	0.80	0.94	4	SoMo B	0.97	0.97	0.98
5	dAtt A		**0.59**	**0.53**	0.95	5	dAtt C	**0.52**	**0.72**	0.99
6	dAtt B		**0.58**	**0.59**	0.96	6	dAtt A	**0.77**	0.92	0.99
7	dAtt C		**0.74**	**0.67**	0.96	7	vAtt A	0.92	0.95	0.98
8	vAtt A		**0.58**	**0.52**	0.96	8	Lim B	**0.66**	0.87	0.93
9	Lim A		**0.49**	0.80	0.94	9	Lim A	0.81	0.91	0.95
10	FPN A		**0.75**	**0.75**	0.97	10	FP B	0.94	0.96	0.98
11	FPN B		0.93	0.95	0.96	11	FP A	0.96	0.98	0.98
12	DMN A		0.88	0.93	0.96	12	DMN A	0.97	0.93	0.98
13	DMN B		0.80	0.87	0.95	13	DMN B	0.87	**0.57**	0.98
14	DMN C		**0.77**	0.91	0.96	14	DMN D	0.97	**0.45**	0.98
	Mean		**73%**	79%	96%			87%	87%	98%

Supplementary Table 2: Stability of transfer learning for decomposition methods. Absolute mean (over participants) correlation between RSN from the Train and Test dataset, when using transfer learning approach A1, A2 and A3. Test RSNs were matched according to highest correlation, while keeping a 1:1 matching constraint. RSN with a correlation < 0.75 are marked with bold, and these are visualized in Supplementary Figure 2. Vis: Visual, SoMo: Somatomotor, dATT: dorsal attention, vATT: ventral attention (also referred to as salience network), Lim: limbic, FPN: fronto parietal network, DMN: default mode network.

Supplementary Figure 2: Stability of transfer learning for decomposition methods. Visual QC of networks with a spatial correlation < 0.75 between the discovery (D1) and validation (D2) dataset, when using transfer learning approach A1 and A2 (networks that are listed with bold in Supplementary table 2). For visualization the ICA RSNs were cut off at |Z| > 1, and for MSAA networks include voxels with >10% fractional contribution. Vis: Visual, SoMo: Somatomotor, dATT: dorsal attention, vATT: ventral attention (also referred to as salience network), FPN: frontoparietal network, DMN: default mode network.

Supplementary Table 3: Classification performance of individual RSNs

ICA									MSAA							
Train		A1		A2		A3			Train		A1		A2		A3	
RSN	AUC	RSN	AUC	RSN	AUC	RSN	AUC		RSN	AUC	RSN	AUC	RSN	AUC	RSN	AUC
'SoMo A'	0.81	'dATT C'	0.78	'dATT C'	0.74	'dATT C'	0.78		'SoMo A'	0.81	'SoMo B'	0.79	'SoMo B'	0.78	'SoMo B'	0.79
'vATT A'	0.80	'FP B'	0.71	'SoMo A'	0.69	'FP B'	0.71		'Vis B'	0.79	'Vis B'	0.71	'dAtt C'	0.76	'Vis B'	0.74
'Lim A'	0.80	'Vis A'	0.68	'FP B'	0.69	'Vis A'	0.68		'Lim A'	0.77	'SoMo A'	0.69	'Vis B'	0.76	'FP A'	0.69
'SoMo B'	0.78	'Lim A'	0.68	'Vis A'	0.69	'Lim A'	0.68		'Lim B'	0.77	'FP A'	0.68	'SoMo A'	0.69	'vATT A'	0.69
'dATT C'	0.76	'SoMo A'	0.67	'DMN A'	0.66	'SoMo A'	0.67		'SoMo B'	0.77	'FP B'	0.65	'FP B'	0.67	'dAtt C'	0.68
'Vis B'	0.76	'DMN A'	0.64	'Lim A'	0.65	'DMN A'	0.64		'Vis A'	0.75	'vATT A'	0.65	'FP A'	0.67	'SoMo A'	0.67
'DMN B'	0.74	'FP A'	0.64	'FP A'	0.65	'FP A'	0.64		'DMN D'	0.75	'DMN D'	0.64	'vATT A'	0.66	'DMN D'	0.67
'dATT B'	0.73	'Vis B'	0.63	'Vis B'	0.63	'Vis B'	0.63		'vATT A'	0.75	'DMN B'	0.62	'Lim A'	0.63	'FP B'	0.65
'dAtt A'	0.73	'dAtt A'	0.63	'dATT B'	0.62	'dAtt A'	0.63		'dAtt C'	0.73	'DMN A'	0.61	'DMN A'	0.61	'Lim A'	0.65
'Vis A'	0.73	'SoMo B'	0.61	'vATT A'	0.60	'SoMo B'	0.61		'FP B'	0.71	'Vis A'	0.59	'DMN D'	0.61	'DMN B'	0.62
'FP B'	0.73	'DMN B'	0.60	'DMN C'	0.59	'DMN B'	0.60		'FP A'	0.69	'Lim A'	0.56	'Vis A'	0.58	'DMN A'	0.61
'FP A'	0.72	'DMN C'	0.60	'dAtt A'	0.59	'DMN C'	0.60		'DMN A'	0.68	'dATT A'	0.53	'dATT A'	0.58	'Vis A'	0.59
'DMN A'	0.70	'vATT A'	0.59	'DMN B'	0.57	'vATT A'	0.59		'DMN B'	0.66	'Lim B'	0.52	'DMN B'	0.52	'dATT A'	0.56
'DMN C'	0.67	'dATT B'	0.53	'SoMo B'	0.56	'dATT B'	0.53		'dATT A'	0.65	'dAtt C'	0.51	'Lim B'	0.52	'Lim B'	0.55

Supplementary Table 3. Classification performance (CP) of individual networks. CP measured using the ROC AUC for the individual resting state networks (RSN) for both multi-subject archetypal analysis (MSAA) and independent component analysis (ICA). AUC listed AUC listed for the training dataset (using 10-fold CV) and on the Test dataset (data from two independent sites), using all three transfer learning approaches, A1, A2 and A3. Vis: Visual, SoMo: Somatomotor, dATT: dorsal attention, vATT: ventral attention (also referred to as salience network), Lim: limbic, FP: fronto parietal network , DMN: default mode network.

Interpretation: overall all RSN yield significant classification (accessed using maximum permutation statistics with 1000 permutations) on the training dataset, and most on the Test dataset (those not significant marked with grey). For the Train dataset, both ICA and MSAA found the highest CP was obtained with the SenMo A network, whereas this network was still significant on the Test dataset, it was no longer the highest. The best Test CP for ICA is the dATT C and for MSAA the SenMo C. Overall there is a high alignment between the order of the CP between the transfer learning approaches A1-A3.

Supplementary Figure 3: Weight vector for diagnosis classification with ROI features

A) Absolute Weightmap

B) Mean weight (z-scored)

Supplementary Figure 3: Weigh vector for resting state networks (RSN). Panel A shows the weightmap for the disease classification using the Haufe transform[2]. Significance was accessed using 1000 random permutations, i.e. the weightmap we only kept ROI connectivities that had significantly higher weight when classifying based on the true, compared to random permutations of the disease labels (p<0.05, resulted in 2684 significant weights). For visualization, the absolute values of the colormap is shown. Panel B shows the mean Z-score on non-zero weights for each ROI RSN. SenMo: Somatomotor, dATT: dorsal attention, vATT: ventral attention (also referred to as salience network), FPN: fronto parietal network, DMN: default mode network. Colors of text correspond to colors on panel A.

Interpretation: Overall within (yellow) RSN connectivities have higher mean Z-scores than between (blue), though there is not much difference. Furthermore, the mean z-score of the RSNs are similar, with highest weight for the FP between network connecitivies.

Supplementary Figure 4: Connectome difference between patients with Schizophrenia and healthy controls
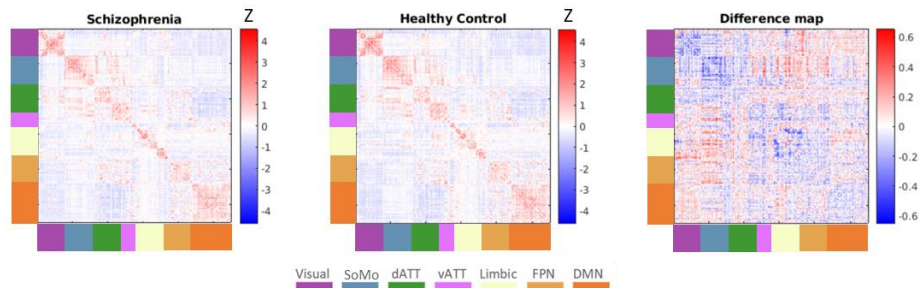
Supplementary Figure 4: Connectome for Schizophrenia, Healthy controls and difference map
FC matrix for patients with schizophrenia (left), healthy controls (middle) and difference map (left) between the two groups. SoMo: Somatomotor, dATT: dorsal attention, vATT: ventral attention (also referred to as salience network), FPN: fronto parietal network, DMN: default mode network. Colors of text correspond to colors on panel A.

Interpretation: The difference maps shows that patients with SZ Have a consistent decreased activity along the diagonal, i.e., hypoconnectivity within each RSN, whereas the between RSN connectivity is a mixture of both hypo and hyper connectivity.

Supplementary Table 4: Prediction performance with different ML models

PANSS total prediction using ROI connectivities and different ML models	Rho
Linear regression	0.30
SVR with linear kernel*	0.30
GPR with linear kernel	0.31
GPR with RBF Kernel	0.31
GPR with Matern kernel	0.31

Supplementary table 4: Prediction performance of different machine learning (ML) models. Prediction performance measured using Pearson's correlation coefficient (r) on dataset D1a. The aim was to compare different models regarding to their stability (from Train to Test) and performance. All models were implemented using Pythons sklearn package[1]. On the Train dataset, 10-fold cross validation (CV) was used with 20 random splits. Performance was calculated on the mean predicted PANSS total score across splits. Specifications for models:

- Linear regression ('LinearRegression')
- Linear support vector regression ('svm.SVR'), kernel range: $10^{\wedge}(-3,3)$
- Gaussian process regression (GRP) ('GaussianProcessRegressor') with 3 kernels:
    - Linear: kernel= C(1.0, (1e-3, 1e7)) + DotProduct(100,(1e-3, 1e7))
    - Radial basis function(RBF) : kernel = C(1.0, (1e-3, 1e7)) * RBF(100,(1e-3, 1e7))
    - Matern: kernel = C(1.0, (1e-3, 1e7)) * Matern(100, (1e-3, 1e7), nu = 1.5)

Supplementary Table 5: PANSS total prediction performance ROI and ensemble decomposition models

Prediction performance		D1a				D2a			
		rho	p	r	p	rho	p	r	p
Total	Parcellation	0.30	0.001	0.36	0.001	0.06	0.33	0.09	0.322
	ICA	0.31	0.001	0.39	0.001	0.27	0.01	0.28	0.006
	MSAA*	0.33	0.001	0.38	0.001	0.12	0.10	0.05	0.238

Supplementary Table5: Prediction performance of PANSS total: Prediction performance for the parcellation based and ensemble decomposition methods. Values also shown in Figure 6 in the main manuscript. Performance measured by Spearman's rank correlation coefficient (Rho) and Pearson's correlation coefficient (r) and significance is accessed using 1000 permutations of the PANSS scores.

Supplementary Table 6: PANSS total prediction of individual RSNs

| | PANSS total of ICA and MSAA | | | | | | | |
| | D1a | | | | D2a | | | |
	Rho	p	r	P	Rho	P_{UC}*	r	P_{UC}*
ICA								
dAtt A	0.31	0.03	0.30	0.04	0.19	0.08	0.22	0.06
vAtt A	**0.27**	**0.05**	**0.33**	**0.01**	**0.22**	**0.04**	**0.22**	**0.04**
DMN B	0.34	0.01	0.41	0.00	0.12	0.16	0.13	0.12
DMN C	0.29	0.05	0.32	0.02	-0.14	0.80	-0.17	0.83
MSAA								
Vis A	0.29	0.03	0.29	0.03	0.08	0.28	0.13	0.17
Vis B	0.27	0.04	0.27	0.05	-0.08	0.68	-0.13	0.79
vAtt A	**0.33**	**0.01**	**0.32**	**0.01**	**0.12**	**0.16**	**0.17**	**0.10**
Lim B	0.33	0.01	0.36	0.00	-0.20	0.89	-0.11	0.70
FP B	0.27	0.05	0.31	0.01	-0.24	0.90	-0.13	0.76
FP A	0.27	0.05	0.29	0.03	0.04	0.32	0.08	0.26

Supplementary table 6: PANSS total prediction for RSNs that are significant on both datasets. Prediction performance measured by Spearman's rank coefficient of correlation (Rho) and Pearson's correlation coefficient (r) for all RSNs that obtained significant prediction on the discovery dataset (D1a). Significance is accessed using 1000 permutations of the PANSS scores. Both ICA and MSAA found the best performance to be for the vATT RSN (marked with bold), which are illustrated in Figure 7 of the main paper.

Supplementary Table 7: Pearson's correlation coefficient for the subscales

| | PANSS prediction of subscales (r) | | | | | |
| | Positive | | Negative | | Generalized | |
	D1a	D2a	D1a	D2a	D1a	D2a
Parcellation	0.22	0.15	0.38	0.07	0.3	0.09
ICA	0.18	0.38	0.35	0.08	0.28	0.22
MSAA	0.11	0.14	0.38	-0.16	0.33	0.11

Supplementary table 7: Pearson's correlation coefficient for ensemble prediction of PANSS subscales Prediction performance measured by Pearson's correlation coefficient (r) for the parcellation based and ensemble decomposition methods (ICA and MSAA) for the three PANSS subscales positive, negative and generalized.

Supplementary Table 8: PANSS dimensions prediction of individual RSNs (ICA)

	Positive		Negative		Generalized	
	D1a	D2a	D1a	D2a	D1a	D2a
Vis A	0.05		0.25		-0.01	
Vis B	0.06		0.17		0.25	
SoMo A	-0.01		0.25	0.07	-0.02	
SoMo B	0.06		0.21		0.11	
dAtt A	0.13		0.22		0.27	
dAtt B	0.00		0.24		0.03	
dAtt C	0.05		0.19		0.18	
vAtt A	0.27	0.05	0.16		0.19	
Lim A	-0.01		0.12		0.19	
FP A	0.04		-0.04		-0.17	
FP B	0.12		0.21		0.33	
DMN A	0.20		0.24		0.15	
DMN B	0.11		0.41	0.01	0.26	0.03
DMN C	0.27	0.14	0.35	-0.11	0.29	-0.04

Supplementary table 8: PANSS subscale prediction for RSNs for ICA. Prediction performance measured by Spearman's rank coefficient of correlation (Rho) for all RSNs in the discovery dataset (D1a) and performance on the independent test dataset D2a, for those RSN that were significant on D1a. Significance is accessed using 1000 permutations of the PANSS scores.

Supplementary Table 9: PANSS dimensions prediction of individual RSNs (MSAA)

| | MSAA (Rho) | | | | | |
| | Positive | | Negative | | Generalized | |
	D1	D2	D1	D2	D1	D2
Vis A	0.13		0.28	-0.13	0.23	
Vis B	0.13		0.25	-0.07	0.27	-0.04
SoMo A	-0.13		0.25	0.04	0.04	
SoMo B	0.02		0.26	0.14	0.06	
dAtt C	0.08		0.16		0.17	
dAtt A	-0.07		0.23	-0.14	0.16	
vAtt A	0.10		0.30	-0.02	0.23	0.10
Lim B	0.06		0.35	-0.05	0.27	-0.15
Lim A	0.06		0.23		0.19	
FP B	0.05		0.33	-0.24	0.18	
FP A	0.16		0.19		0.22	
DMN A	-0.03		0.31	-0.03	0.17	
DMN B	0.11		0.22		0.23	
DMN D	0.10		0.22		0.14	

Supplementary table 8: PANSS subscale prediction for RSNs for MSAA. Prediction performance measured by Spearman's rank coefficient of correlation (Rho) for all RSNs in the discovery dataset (D1a) and performance on the independent test dataset D2a, for those RSN that were significant on D1a. Significance is accessed using 1000 permutations of the PANSS scores.

Supplementary Table 10: Acquisition parameters for resting state fMRI data across 10 protocols

	Discovery dataset D1								Validation data D2	
Site number	1	2	3	4	5	6	7	8	9	10
Site acronym	HKH	COI	KTT	UTO	ATV	ATT	CIN	COBRE	SUW	KUT
n participants	29	123	121	132	39	13	39	133	120	203
n HC	29	123	75	96	39	13	39	72	101	159
n SZ	0	0	46	36	0	0	0	61	19	44
PANSS scores avaliable	NA	NA	yes	yes	Na	Na	Na	yes	NA	yes

MRI aqucition parameters

MRI Scanner	SIEMENS Symp	SIEMENS Verio	SIEMENS Trio	GE Discov	SIEMENS Verio	SIEMENS TimTrio	SIEMENS TimTrio	SIEMENS TimTrio	SIEMENS Verio	SIEMENS TimTrio
Magnitic field strength	3T	3T	3T	3T	3T	3T	3T	3T	3T	3T
Number of channels per coil	head-12ch	head-12ch	head- 8ch	head-24ch	head-12ch	head-12ch	head-12ch	head-12ch	head-12ch	head-32ch
TR (s)	2.7	2.5	2	2.5	2.5	2.5	2.5	2	2.5	2.5
TE (ms)	31	30	30	30	30	30	30	29	30	30
Flip angel (deg)	90	80	90	80	80	80	80	75	80	80
Phase encoding	AP	AP	AP	PA	PA	PA	AP		PA	PA
Matrix	64 x 64	64 x 64	64 x 48	64 x 64	64 x 64	64 x 64	64 x 64	64 x 64	64 x 64	64 x 64
Field of view (mm)	192	212	256 x192	212	212	212	212	240	212	212 x 212
In-plane resolution (mm)	3.0 x 3.0	3.3 x 3.3	4.0 x 4.0	3.3	3.3 x 3.3	3.3 x 3.3	3.3 x 3.3	3.75 x 3.75	3.3 x 3.3	3.3125 X 3.3125
Slice thickness (mm)	3	3.2	4	3.2	3.2	3.2	3.2	3.5	3.2	3.2
Slice gap (mm)	0	0.8	0	0.8	0.8	0.8	0.8	1.05	0.8	0.8
Number of slices	38	40	30	40	39	39 or 40	41	33	40	40
Slice acquisition order	Ascending (IL)	NA	Ascending (IL)	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending
Number of volumes	107 + 5 (dummy)	240 + 4 (dummy)	182	240 + 4 (dummy)	240 + 4 (dummy)	240 + 4 (dummy)	240 + 4 (dummy)	149 + 1 (dummy)	240 + 4 (dummy)	240 + 4 (dummy)
Total scan time	~5 min	10 min	6 min	10 min	10 min	10 mins	10 min	~ 5 min	10min	10 min
Eye closed/fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
Field map:										
E Echo spacing	NA	0.0005	NA	0.00029	0.00049	0.00049	NA	NA	0.0005	0.00056
Echo time 1	NA	0.00492	NA	0.0049	0.00492	0.00492	NA	NA	0.00492	0.00492
Echo time 2	NA	0.00738	NA	0.0074	0.00738	0.00738	NA	NA	0.00738	0.00738
Blipdir	NA	j-	NA	j	j	j	NA	NA	j	j

Supplementary table 10: Number of participants and MRI acquisition parameters for each site. The top part of this table indicates the site acronym (HKH, Hiroshima Kajikawa Hospital; COI, Center of Innovation at Hiroshima university; KTT, Kyoto University (Trio); UTO, University of Tokyo Hospital; ATT, Brain Activity Imaging Center ATR-Promotions Inc., Kyoto (Trio); ATV, Brain Activity Imaging Center ATR-Promotions Inc., Kyoto (Verio); CIN, Center for Information and Neural Networks; COBRE, The Center for Biomedical Research Excellence ; SWA, Showa university; KUT, Kyoto University (TimTrio)) and number of participants for each site. The bottom indicates the MRI acquisition parameters. We chose to exclude all data from two DecNef acquisition sites (HUH and HRH) since our visual quality control showed that the EPI images from these sites had many artifacts and a low signal to noise ratio, after the five first (excitation) volumes were excluded. More information about the sites from the DecNef database (all apart from COBRE), can be found in Tanaka et al. [4].

Supplementary figure 4 Yeo 7-parcellaiton network
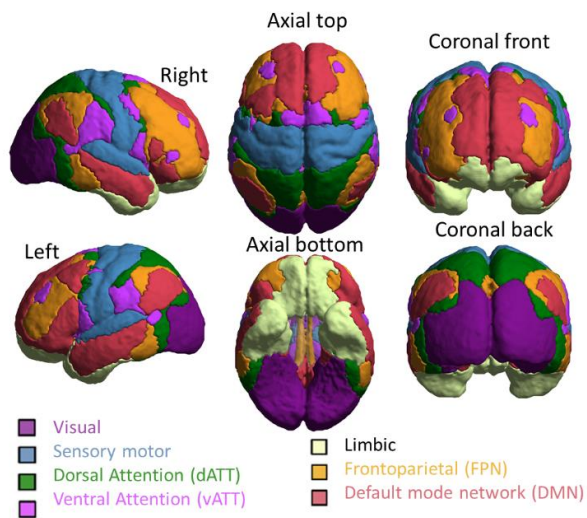
Supplementary Figure 5: Yeo 7-network parcellation from atlas. Illustration of the 7 resting state networks (RSN) from the Yeo parcellation, as downloaded from: CorticalParcellation_Yeo2011 - Free Surfer Wiki (harvard.edu) . The parcellation map was mapped on the cortical surface by using the BrainNet Viewer package (http://www.nitrc.org/projects/bnv).

References:

1. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**(null): p. 2825–2830.
2. Haufe, S., et al., *On the interpretation of weight vectors of linear models in multivariate neuroimaging.* NeuroImage, 2014. **87**: p. 96-110.
3. Yeo, B.T., et al., *The organization of the human cerebral cortex estimated by intrinsic functional connectivity.* J Neurophysiol, 2011. **106**(3): p. 1125-65.
4. Tanaka, S.C., et al., *A multi-site, multi-disorder resting-state magnetic resonance image database.* Scientific Data, 2021. **8**(1).

PAPER D

Title

On the search for data-driven and reproducible schizophrenia subtypes using resting state fMRI data from multiple sites

Authors

Krohne, Laerke G ;Hansen, Ingeborg H; Madsen, Kristoffer H

Status

In preparation for submission

On the search for data-driven and reproducible schizophrenia subtypes using resting state fMRI data from multiple sites

Lærke Gebser Krohne[1,2], Ingeborg Hansen[1], and Kristoffer H. Madsen[2,3]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
[2] H. Lundbeck A/S, Valby Denmark
[3] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital - Amager and Hvidovre, Copenhagen, Denmark

Abstract. For more than three decades, functional magnetic resonance imaging (fMRI) data has been used to search for objective biomarkers for patients with schizophrenia. However, so far, firm conclusions are still to be made, which has often been attributed to the high internal heterogeneity of the disorder. A promising way to disentangle the heterogeneity is to search for data-driven disease subtypes, which has the potential to find subgroup of patients which have a more homogeneous biological profile.

In this study, we have used an unsupervised multiple co-clustering (MCC) method to identify subtypes on connectivity estimated from a multi-site resting state fMRI dataset. We merged data from two publicly available databases, and split the data into a discovery dataset (data from 3 sites including 143 patients and 143 matched healthy controls (HC)) and an external test dataset (63 patients and 63 matched HC) including data from two independent sites. On the discovery data, we investigated the stability of the clustering towards changes in the dataset and different initializations. Subsequently we searched for cluster solutions with a significant diagnosis association. We further evaluated the clustering results by its subject and feature cluster separability, the included brain features and correlations to clinical manifestations as measured with the Positive and Negative Syndrome Scale (PANSS). Finally, we validated our findings by testing the diagnosis association on the external test data.

We found that the stability of the clustering was highly dependent on variations in the dataset, and even across different initializations we only found a moderate subject clustering stability. Nevertheless, we discovered one clustering solution which had a significant diagnosis association, which reproduced on the external test data. This solution included three subject clusters with a predominance of schizophrenia patients, and a feature cluster that showed a linear trend in the connectivity values for groups with proportions of patients with schizophrenia, which also reproduced on the external dataset. Finally, we did not find any reproducible correlation between the feature clusters and the PANSS scale, indicating the the cluster solution reflects other sources of variability in the data.

To conclude, we see the contribution of this work as an important steps towards establishing the stability, reproducibility and potential of MCC for sub-

typing psychiatric disorders such as schizophrenia, and we hope that it will inform and inspire future work within this important field.

Keywords: fMRI · Subtyping · Schizophrenia · Multiple co-clustering · Stability analysis

1   Introduction

Psychiatric disorders, such as schizophrenia, are typically classified based on diagnostic tools and clinical rating scales are used to measure specific symptoms. A challenge of such symptomatology-based outcomes is that they do not necessarily reflect the underlying mechanism that causes them. In practice, symptoms can arise from different causes, while the same biological cause can also lead to different symptoms and phenotypes [1]. Several initiatives have therefore been established with the aim of uncovering data-driven mechanistic disease definitions to increase our understanding of mental disorders themselves, advance biomarker discovery and to identify best treatments for individual patients [2, 3]. For schizophrenia, there are currently no clinically used biomarkers to assist decision for diagnostic or treatment purposes, but there are substantial research efforts towards this. Functional magnetic resonance imaging (fMRI) is a neuroimaging modality that can be used to map brain activation in the whole brain, including subcortical regions, as well as network interactions, which makes it a promising modality to search for schizophrenia biomarkers [4].

1.1   Functional MRI for disease subtyping

Functional MRI (fMRI) is a noninvasive neuroimaging method which gives an indirect measure of brain activation (most frequently done using the blood-oxygen-level-dependent (BOLD) signal) either during task or at rest. Over the decades, many studies have used fMRI data to determine difference in brain activation between healthy controls and patients with schizophrenia (SZ). Earlier studies have mostly used univariate group analyses [5, 6] or supervised machine learning where the fMRI data is used to make individual predictions [4, 7, 8]. However, due to the high internal heterogeneity of psychiatric disorders such as schizophrenia, it has been argued that it might be more valuable to use fMRI to search for potentially new data-driven disease definitions [1, 4, 9]. Another challenge of traditionally univariate brain mappings analysis and supervised prediction studies is that they heavily rely on the "label" e.g., the diagnosis of a phenotypic measure of interest. Unfortunately, in psychiatric disorders, the use of such labels as a "gold standard" is complex, due to the high heterogeneity within the disorder and other factors such as the test-re-test reliability between raters [10, 11].

An increasing number of studies have therefore started to combine fMRI data with clustering methods (type of unsupervised machine learning) to search for subtypes with a more homogeneous biology [1, 12].

In a systematic literature review from 2021, Miranda et al, grouped earlier fMRI studies that used clustering methods to find subtypes in psychiatric disorders into

top-down and bottom-up approaches [1]. In the *top-up* approaches, fMRI is used to validate subtypes that were found based on symptomatology based outcomes. This is helpful to increase the neurobiological understanding, and to evaluate the states of a patient at a given time. However, these subtypes are likely to yield disease symptomatic states rather than biological entities. On the contrary, *bottom-up* approaches perform clustering directly on the data (e.g. fMRI data or other biomarkers) and thus have the potential to uncover subtypes with a more homogeneous biology, which might be closer to the pathological origin [1]. Finally, Miranda et al. introduced a third category of *polytopic learning methods*, which can serve as an interface between top down and bottom up approaches, that combine datatypes. This can be done either by performing the clustering analysis on both kinds of data (e.g., biomarker and clinical outcome assessments) at the same time, or by relying on multi-model transformations such as canonical correlation analysis [13]. The goal of the latter is to bridge the gap between origin and manifestations of the disease, but these methods are also prone to the risk of overfitting if not applied correctly [1]. Polytopic learning can be beneficial if the data, such as fMRI, are of high dimensionality, where the combined clustering with data from a clinical scale can help to extract neurobiological information about disease related trends, that otherwise have considerable risk of being overlooked [1].

Even though fMRI based subtyping has been a goal for many years, the field is still in an exploratory state, where results and methods are largely divergent and results are rarely replicated [1]. One of the core challenges so far has been the high dimensionality of FC data that has hindered an effective application when using conventional clustering methods. This is both due to the the well-known problem of the curse of dimensional [14], and that FC data has a complex structure such that subjects cluster differently depending on the features of interest [15]. In recent years, new methods have been developed specifically for high dimensional data, which carry a great potential for future research to establish relevant and stable subtypes [16, 17].

Within schizophrenia, we are aware of three bottom-up studies that used fMRI for disease subtyping. The first studies were from Brodersen [18]and Yang et al. [19], which demonstrated that clustering can be used to subtype patients with SZ. However these two studies suffered from a range of methodological challenges [1], and their findings have to the best of our knowledge not yet been replicated. The third study was presented by Tokuda et al. in 2021 [20], where the aim was to establish a common brain network for discriminating between patients with various psychiatric disorders (hereby SZ) and healthy controls. They found significant alterations in a brain networks involving a cerebellum-thalamus-pallidum-temporal circuit, which could differentiate between the different diagnosis. Since this study looked for "disorder differentiation networks" they did not aim to identify subtypes within disorders themselves.

1.2 Multiple co clustering

In our study, we chose to focus on a multiple-co clustering (MCC) subtyping algorithm developed by Tokuda et al. [21], which is specifically developed for high di-

mensional data of different types, and which has shown promising results on fMRI data [12]. The terminology "multiple" here refers to the algorithms ability to split the features into several views, while "co-clustering" denotes that within each view, the algorithm further clusters the data into both subject and feature clusters. Here, the views serve as a "feature selection step", such that several different subject clustering solutions (and thereby potential subtypes) can be found depending of the feature included in each view, which was our primary motivation for choosing the method. Furthermore it has additional advantages including: i) it can include different data types (e.g. both fMRI data and data from clinical scales which can have different distributions) which enables polytopic learning, ii) it can deal with missing data, iii) the number of views, subject and feature clusters are automatically inferred [12,21]. The method is described in more detail in the methods section 2.4.

1.3   Objectives of this study

In our work we aimed to: i) test the stability of the MCC algorithm on a multi-site resting state fMRI dataset, and ii) determine if we can find any subtypes with a significant SZ diagnosis association. Here, we specifically focused on resting state functional connectivity data. We performed the analyses on a multi-site discovery dataset, and furthermore tested the reproducibility of our findings on a external dataset, which included data from two independent test sties.

More specifically, on the **discovery dataset** we:

– Determined the stability across random initializations
– Determined the stability across data splits
– Searched for views with significant diagnosis association

For **views with a significant diagnosis association**, we then

– Evaluated the separability between subject clusters
– Determined the correlation to clinical scales for each feature cluster
– Determined the reproducibility of the diagnosis association on the external data

To the best of our knowledge, this study is the first to apply the MCC method to search for schizophrenia subtypes using multi-site rsMRI data, and to perform a comprehensive evaluation of the MCC algorithms stability using fMRI data.

2   Materials and methods

2.1   Participants and data

We used data from two publicly available datasets: i) the DecNef Project Brain Data Repository (https://bicr-resource.atr.jp/srpbsopen/) [22] and ii) the Center of Biomedical Research Excellence (COBRE) dataset [23]. We split the data into a discovery dataset (D1) that contained data from 8 sites (COBRE + 7 DecNef sites) and a test dataset (D2) containing data from two independent sites (both from the DecNef database) which we used to test the reproducibility of our findings. To avoid that age

and gender effects leads to apparent differences between control and schizophrenia related groups, we constructed a balanced dataset using the R package MATCHIT [24]. As the databases included much more data from healthy controls, each patient with SZ was matched to a control with nearest neighbor matching based on propensity scores (age and gender) and exact matching on site. The quality of the matches were assessed through the balance of the covariates (age and sex) before and after matching (quantitatively and using diagnostic quantile-quantile plot (QQ) plots), and visual inspection of the propensity score distributions. Participant demographics are given in Table 1.

	Discovery (D1)		Test (D2)	
	HC	SZ	HC	SZ
n_participant	143	143	63	63
Gender ($\male$, $\female$)	101/42	100/43	37/26	35/28
n_sites	3	3	2	2
Age (μ, σ)	35 ±10	36 ±12	42 ±11	42 ±10

Table 1. Participant demographics. Number of participants (healthy control(HC) and patients with schizophrenia (SZ)), gender, number of sites and age for the participants included in the discovery (D1) and independent validation Test set (D2).

For all participants we used a structural T1 weighted MRI scan and functional MRI data (5-10 min resting state fMRI recorded with open eyes). This study was approved by the Institutional Ethical Review Board at the Technical University of Denmark, department for applied Mathematics and Computer Science (COMP-IRB-2022-03).

2.2   Preprocessing

The MRI data was preprocessed using the fMRIprep v. 20.2.6 pipeline [25] with standard settings for slice timing correction, realignment, registration between T1 and fMRI, segmentation, and spatial normalization to standard (MNI) space [25]. We then used a high pass filter with a cut off frequency of 0.008 Hz, and smoothed the images with a 6mm FWHM isotropic Gaussian filter. For some sites, we had scans assessing the B_0 field inhomogeniety available, for which we estimated and applied a voxel displacement map based on the effective echo spacing and phase encoding direction. For the remaining data (sites where no B_0 map was available), we used the "Fieldmap-less" distortion correction by matching the anatomical features from the T1-weighted scan [25]. We regressed out the mean signal of nuisance compartments (cerebrospinal fluid, white matter and global mean signal) and 24 motion parameters [26]. Finally, we explicitly modelled volumes where the framewise displacement was higher than 1mm, to effectively remove their influence on the results.

2.3   Parcellation based connectivity analysis

To extract the functional connectivity (FC) of the rsfMRI data, we used a parcellation based connectivity approach . First we extracted the time series of the BOLD signal from an atlas. Secondly the connectivity matrix was calculated using Pearson's correlation coefficient with a subsequent Fisher transform and z scoring. For the brain parcellation we used the *Allen atlas,* which is an ICA based atlas based on resting state data from 603 healthy controls [27]. We implemented the atlas using the MASKER function from NILEARN (version 0.9.0) in PYTHON (version 3.9). This atlas includes 28 components which are assigned to one of the following resting state networks (RSNs): basal ganglia, auditory, sensory motor, visual, default mode, attentional and frontal, as specified by Allen et al. [27].
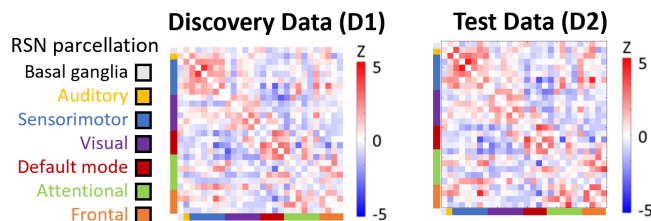
Fig. 1. Functional connectivity (FC) matrix by Allen parellation. Mean (across participants) FC matrix for the discovery (D1) and test (D2) datasets. The colorcode on the axis indicates the corresponding resting state network (RSN) for each brain parcel. Overall, we found a similar connectivity pattern across the two datasets, with a dominance of positive connectivity within the RSNs (along the diagonal) whereas the connectivity between RSN were more mixed.

2.4   Multiple-co clustering algorithm

The multiple co-clustering (MCC) method presented by Tokuda et al in 2017 [21], is a polytopic learning method, that can deal with data types that follow different distributions. As described in the introduction the key idea is that the algorithm optimally partitions the features into several groups (called views) in which subject and feature clustering is performed separately. The algorithm simultaneously partitions the data in the following way:

- Partition features into several views (works as feature selection for different subject cluster solutions)
- Further partitions each view into feature clusters (bundling similar features)
- Partition participants into subject clusters

The clustering is based on non-parametric Bayesian Mixture models, where mixing of several types of distributions are allowed. The current implementation of the

algorithm allows for Gaussian, Bernoulli and Poisson distributions depending on the underlying data types. The number of views, feature and subject clusters are automatically inferred based on a Dirichlet process prior. The generative models for the view, feature and subject clustering can be found in earlier descriptions of the model by Tokuda et al. [12,21], where the feature clusters $\mathbf{Y}$ and subject clusters $\mathbf{Z}$ indicators are given by a multinominal distribution (generated by hierarchical stick breaking process) such that:

$$Y_{j...}^{(m)} \, \mathrm{Mul}(\cdot | \tau^{(m)}) \tag{1}$$

denotes the view and feature cluster membership vector for feature j of the distribution family m. And the subject cluster membership indicator for each subject i is given by

$$Z_{i...}^{(m)} \, \mathrm{Mul}(\cdot | \eta_v) \tag{2}$$

The model assumes that each instance of the data matrix $X_{i,j}$ independently follows a specified distribution conditional on $\mathbf{Y}$ and $\mathbf{Z}$. The log-likelihood of the model is

$$\log p(\mathbf{X}|\mathbf{Y},\mathbf{Z},\Theta) = \sum_{m,v,g,k,j,i} \mathbb{I}(Y_{j,v,g}^{(m)} = 1)\mathbb{I}(Z_{i,v,k} = 1) \log p(X_{i,j}^{(m)} | \theta_{v,g,k}^{(m)}). \tag{3}$$

where $\theta_{g,k}^{m}$ are the parameters of the distribution family m for feature cluster g, subject cluster k and view v.

In our study, most variables were from the functional connectivity matrix (Gaussian distribution), but we also included a few additional features: age (Gaussian distribution) as well as gender and site allocation of each participant (categorical which were modelled with a Bernoulli distribution). The conjugate priors for the Gaussian distributed features were given by the normal inverse gamma distribution, which is defined by four parameters, μ_0 , σ_0^2, λ_0 and γ_0 and the Bernoulli distributed features are given by a single parameter β_0 (supposing a symmetric Dirichlet distribution) [21]. The priors, expectation of the log-likelihood and update equations are described in the supplementary section of Tokuda et al. in 2017 [21] and the source code of the algorithm which is publicly available at Github (https://github.com/tomokitokuda/Multiple-Co-clustering).

2.5   Missing data and non-imaging features

The MCC method can handle missing values if they occur at random, such that missing entries are considered as stochastic parameters, which in practice is done by marginalizing across the missing entries such that it ignores these when it updates the hyper-parameters, [21]. To evaluate the potential of polytopic learning with the MCC algorithm, we would have preferred to include both functional connectivity, demographics and clinical scales data. However, since the latter was only available for patients with schizophrenia, and not our healthy control participants, we opted against including data from clinical scales in our subtyping analysis such that we did not violate the assumption of missing at random.

In this study, we therefore included the following features: 1) 378 FC connectivity features from the Allen atlas 2) age and 3) five binary features indicating the gender,

handedness and three binary labels to indicate at which site the data of the participant was acquired (the discovery dataset was from three different sites as specified in section 2.1).

Different initializations The clustering solution of the MCC method will depend on the initialization, which is given by a random configuration of views and clusters. As in earlier studies, we used a "5-step heuristic" to balance how well the model described the data (assessed by log likelihood) and the stability across initializations [14, 20]. The clustering stability was assessed by using the adjusted rand index (ARI), which is a measure of the similarity between two clustering (adjusted to correct for chance by using the permutation model for clustering [28].) Our primary stability measure was the "feature to view" clustering stability (ARI_{view}), which measures the similarity of what features are assigned to what views (but does not evaluate the subject and feature clustering within the view).

The "5-step heuristic" includes the following steps:

1. Run algorithm with 1000 different initializations
2. Find the 10 models with the highest log likelihood (equation 3)
3. Calculate the "feature to view" membership stability (ARI_{view}) between the top ten models
4. Identify the pair with highest ARI_{view} (this will be referred to as "top-two-pair" for the remaining paper)
5. The final model, used for subsequent analysis, is then the one in the top-two-pair which had the highest log likelihood

2.6   Stability analysis

We two stability analyses to evaluate the stability of the algorithm with regards to 1) different initializations and 2) data splits as summarized in Figure 2.
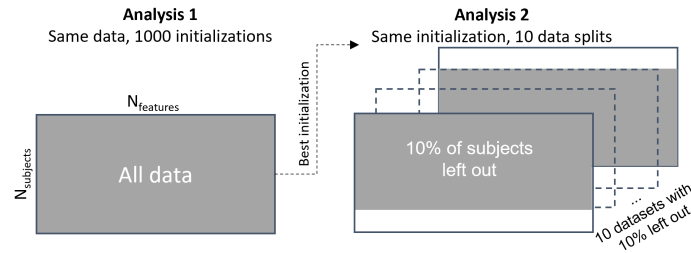
Fig. 2. Stability analysis of subtyping. In stability analysis 1 (left) the MCC algorithm is run 1000 times using the same dataset and different initializations. In analysis 2 (right), 10 sub-datasets were created, where 10% of the data (28 participants) were left out for each dataset. In this way, we tested the stability of the algorithm to the expected variability in the data.

Stability analysis 1: initializations

To measure the stability across different initializations, we used the same "5-step heuristic" as in the earlier publications by Tokuda et. al [14, 20], and we assessed stability in the following way:

– View membership stability was measured using the ARI_{view} (mean across all top 10 models from step 3)
– To get a better understanding of the similarity across runs, we also calculated the subject ($ARI_{subject}$) and feature ($ARI_{feature}$) cluster similarity for the top-two-pair with highest ARI_{view} (step 4). We performed this analysis for each view separately, where we used a Procrustes alignment procedure to match views across the two runs with regards to maximal feature overlap.

We refer to each of the 1000 initializations as "runs", where each run has a new random initialization that determines the initial assignment of the data into the views as well as feature and subject clusters. For each run the algorithm performed 1000 iterations (this was a pragmatic choice since 1000 iterations already resulted in a run time 10 hours per run). Whereas earlier studies have used the " 5-step heuristic" to choose the "best" solution of 1000 initializations, they have not reported the ARI for the top 10 models nor top-two-pair. To the best of our knowledge there are also no other studies that have reported stability measure of the MCC algorithm on neuroimaging data (the only formal stability analysis we have found reported was on simulated data in the original publication of the method in 2017 [21]).

Stability analysis 2: data splits

Whereas the first stability analysis determined the clustering similarity when using different initializations on the same dataset, the aim of this analysis was to investigate how robust the clustering similarity was towards leaving out a part of the data. For this analysis we kept the initialization constant (selected the best initialization from stability analysis 1) at then reran the clustering on 10 "sub-datasets", where we left out 10% of the data (28 subjects) for each dataset as illustrated in Figure 2. To the best of our knowledge this is the first time that the stability of the MCC algorithm has been investigated across data splits.

2.7   Views with diagnosis association

To search for potential subtypes that are related to the schizophrenia diagnosis, we first identified views with a significant diagnosis association by using Pearson's χ^2 test for contingency tables, where we evaluated the association between the subject-cluster label and diagnosis label. The Pearson's χ^2 test statistic is used to test the independence between the rows (R) and columns (C) of the contingency table, where independence refers to knowing the value of the row variable (here subject-cluster label as estimated by the MCC clustering) does not change the probabilities of the column variable (diagnosis label), and vice versa. The Pearson's χ^2 test statistics follows an asymptotic χ^2 distribution with (R-1)(C-1) degrees of freedom, which is cal-

culated as:

$$\chi^2 = \sum_i \sum_j \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}. \tag{4}$$

Where $O_{i,j}$ is the observed count for the ith row ($i = 1$ to R) and j^{th} column ($j = 1$ to C) in the contingency table. And $E_{i,j}$ is the expected counts when assuming independence (null hypothesis of the test) which is calculated as $E_{i,j} = \frac{n_{i.} n_{.j}}{N}$ where $n_i.$ and $n_.j$ are the row and column marginal totals, and N is the total number of counts in the table.

We assessed significance by using random permutation testing to obtain an empirical null of χ^2 by creating 1000 random permutations of the diagnosis label [29]. We corrected for multiple comparisons using maximum permutation statistics, i.e., we created an empirical null distribution by considering only the most significant effect over the entire set (all views for that run). This controls the family-wise error over the set.

Sorting of subject and feature clusters
To ease the visual interpretation of the views, we sorted the views such that subject clusters were sorted in ascending order of proportions of patients with SZ, (i.e., subject clusters with highest proportion of HCs are on the top), and furthermore within each subject cluster healthy participants are displayed first, and patients with SZ are indicated with a black hyphen . Feature clusters were sorted by descending cluster size. Please note that the diagnosis label was only used to sort the subject clusters, but not included as a feature in the MCC at any point.

Separability between subject clusters
For views with a significant diagnosis association, we looked further into the different subject clusters to search for potential schziophrenia-related subtypes. As in the earlier neuroimaging applications of the MCC clustering, we used the Cohen's D to measure the separability between two neighboring subject clusters (neighboring since the clusters are sorted according to the proportion of patients with SZ). Cohen's D (CD) is a commonly used measure of effect sizes for differences between two distribution, and we use the following descriptors for magnitudes: < 0.5: small, 0.5-0.8: moderation, > 0.8 large, > 1.2: very large [30]. Whereas earlier studies evaluated CD for each view by taking the mean CD of all neighboring subject clusters and feature clusters included within a view, we report the CD for each feature cluster and subject-cluster pair separately. Finally, to gain more insight about the difference between all the subject clusters (and not only the neighboring) we also showed the histogram for each subject-feature cluster, as well as the probability density function (PDF) using the posterior hyper-parameters from the MCC algorithm. The PDF for the normal inverse gamma distribution is given as

$$f(x, \sigma^2 | \mu, \lambda, \alpha, \beta) = \frac{\sqrt{\lambda}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{2\beta + \lambda(x-\mu)^2}{2\sigma^2}\right). \tag{5}$$

Here μ and λ are directly given by the hyper-parameters and $\alpha = \gamma_0/2$ and $\beta = \gamma_0\sigma_0^2/2$ are given as specified in the supplementary material of the earlier MCC publications [12, 21], and Γ is the gamma distribution.

Relation to clinical scales

For views with a significant diagnosis association, we also investigated if any of the subject-feature clusters were related to the clinical differentiation given by the PANSS score. For this analysis we only included patients (no HC) with an available PANSS score (4 patients did not have these available). More specifically, for each patient we calculated the mean FC for each feature cluster, and correlated it to the PANSS score for that patient. Since the PANSS scale is a summation of categorical subitems (and thus not a continuous measure), we used the Spearman's rank coefficient of correlations (Rho), which is a non-parametric measure of correlation utilizing ranks [31], and has less sensitivity to outliers (patients with very high or low PANSS scores). We used the total PANSS score as a measure of symptom severity, and furthermore the three subscales PANSS positive, negative and generalized, to determine if any of feature clusters were specifically related to any of dimensions given by these three PANSS subscales.

2.8   Reproducibility on independent test dataset

Whereas all previous steps were performed on the discovery dataset, we kept approximately 30% of the data as an independent test dataset to evaluate the reproducibility of our findings. We aimed to determine if views that had a significant diagnosis association on the discovery dataset (View$_{\text{diagnosis, D1}}$), also showed a significant diagnosis associated subject clustering on the test dataset (D2). This was done using the following **four steps**:

1. **Test data, View$_{\text{D2}}$:** select features that were included in View$_{\text{diagnosis, D1}}$ and keep the feature cluster solution fixed, such that that each feature is assigned to the same feature cluster as in the discovery dataset
2. **Model parameters** θ_{D1}**:** extract the hyper-parameters from each subject-feature cluster given the MCC solution of the discovery dataset.
3. **Expectation** $\mathbb{E}_{q(\theta)}$**:** calculate the expectation of the conditional log likelihood for each possible subject-feature cluster combination using θ_{D1} on the test dataset View$_{\text{D2}}$
4. **Final subject cluster solution:** assign each participant to the subject-cluster that maximizes the log likelihood

Since we found that View$_{\text{diagnosis, D1}}$ only included FC features (no binary feature included in this view, as described in the results section 3.4) the model parameters (step 2), where given by the hyper-parameters from the normal-inverse gamma distribution, μ, λ, γ and σ^2 as described in section 2.4. Given these hyper-parameters,

the expectation of the conditional log likelihood is given by:

$$\mathbb{E}_{q(\theta)}[\log p(X_{i,j}|\theta_{g,k}] =$$

$$-\frac{1}{2}\left\{\frac{(X_{i,j}-\mu_{g,k})^2}{\sigma_{g,k}^2}\frac{1}{\lambda_{g,k}}+\log\left(\sigma_{g,k}^2\right)+\log\left(\frac{\gamma_{g,k}}{2}\right)-\psi(\frac{\gamma_{g,k}}{2})+\log(2\pi)\right\} \tag{6}$$

Where **X** is test dataset View_{D2} (step1). The feature clusters (k) were kept constant (step 1) such that $\mathbb{E}_{q(\theta)}$ was calculated for each subject subject clusters (g) (step 3). The participants were then assigned to the subject-cluster that maximizes the log likelihood given by equation 3.

3   Results

In this section we list and shortly describe the findings of our study. Section 3.1 and 3.2 list the stability analyses and section 3.3 described the views included in the final clustering solution. For this solution, section 3.4, 3.5 and 3.6 include the results of the diagnosis association, subject cluster separability and clinical scale correlation analyses respectively. Finally section 3.7 includes the reproducibility analysis where we tested the estimated subject clustering on an external dataset.

3.1   Stability analysis 1: initializations

In this stability analysis, we kept a constant dataset, and repeated the MCC clustering with 1000 different initializations (also referred to as runs). Our primary stability measure was the ARI_{view}, which measures the view membership similarity across runs. The cost function (log likelihood (LE) as a function of iterations) is shown in Panel A of Figure 3, which shows that some models still had occasional abrupt changes close to the final number of iterations(1000 iterations), indicating that full convergence had not been reached. Panel B of Figure 3 shows the ARI_{view} for all combinations of the top 10 models. The mean ARI_{view} was 0.84 and the top-two-pair (step 4) had an ARI_{view} of 0.99, which indicates that the feature assignment to each view was very similar for these runs.

To get a better understanding of the similarity across runs, we also calculated the subject ($\text{ARI}_{\text{subject}}$) and feature ($\text{ARI}_{\text{feature}}$) cluster similarity for the top-two-pair (run 360 and 965, which will be referred to as Run A and Run B in the remainder of the manuscript). We performed this analysis for each view separately, using a Procrustes alignment procedure to match views across the two runs with respect to maximal feature overlap. Table 2 shows that even though the two runs do not have the same number of views (3 views for run B and 4 views for run A), the included features in each view are very similar, with less that 10 features (out of a total of 378) that were assigned to different views between runs. For view 1 and view 2 we found that the $\text{ARI}_{\text{feature}}$ was high (0.85 and 0.96), whereas the $\text{ARI}_{\text{subject}}$ only was moderate (0.67-0.77).

Greedy matching on views			N features in view		$\text{ARI}_{\text{feature}}$	N feature cluster		$\text{ARI}_{\text{subject}}$	N subject clusters	
Run B	Run A	Feature match	Run B	Run A		Run B	Run A		Run B	Run A
1	1	0.99	331	326	0.85	7	6	0.77	11	10
2	2	0.97	51	52	0.96	4	3	0.67	6	8
3	4	1.00	1	1	1.00	1	1	1.00	1	1
	3			1			1			1

Table 2. Stability analysis 1 of pair with highest ARI_{view}. Subject and feature cluster stability analysis for the pair of runs with highest ARI_{view} (run A and B). Views between the two runs were matched using greedy matching, where 1 indicates that the features in the two views are identical, and 0 means that there is not overlap between features at all. Then the cluster similarity was calculated for the subject clusters ($\text{ARI}_{\text{subject}}$) and feature clusters ($\text{ARI}_{\text{features}}$) for each view separately, and furthermore the table lists the number (N) of feature and subject clusters for each view.
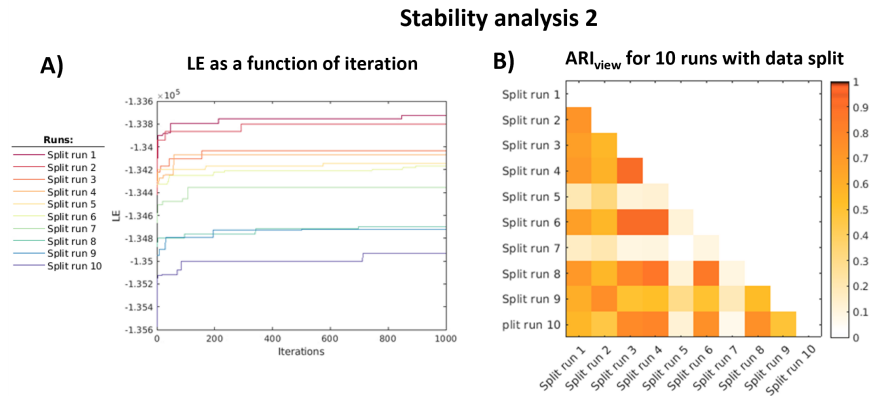
Stability analysis 1
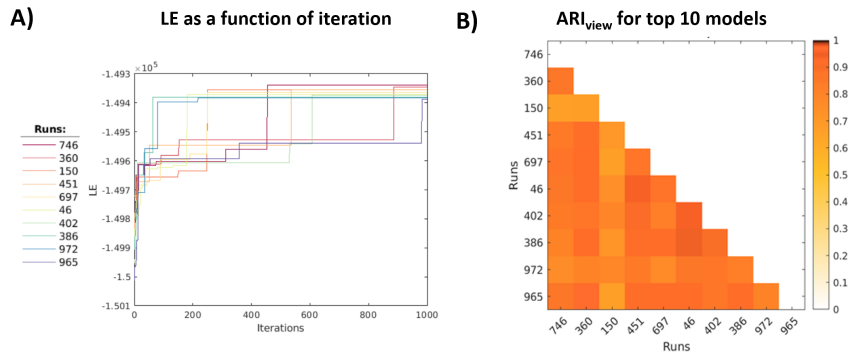
Fig. 3. Stability analysis 1 Panel A shows the log likelihood (LE) as a function of the number of iterations (x-axis) for the ten best runs. Panel B shows the ARI_{view} for all combinations for the top 10 models. The mean ARI_{view} is 0.84, and the top-two-pair with highest ARI_{view} was between run 360 (Run A) and run 965 (Run B). Of these two runs, Run A had the highest log likelihood and will thus be the final model chosen using the 5-step heuristic described in section 2.4

3.2   Stability analysis 2: data splits

The aim of this second stability analysis was to evaluate the stability (ARI_{view}) when leaving out a small part of the data for each run.


Fig. 4. Stability analysis 2 Panel A shows the log likelihood (LE) as a function of the number of iterations (x-axis). Panel B shows the ARI_{view} for all combinations of the ten data split runs. The mean ARI_{view} is 0.48.

Figure 4 shows the result from stability analysis 2, where the cost function in Panel A, shows that the increases in the LE over iterations are small compared to the differences in LE across the runs with left out data. Panel B shows the ARI_{view} for each combination of the runs, which are substantially lower than for stability analysis 1 (Figure 3). These findings showed that leaving out 10 % of the data for each run does influence the stability to a higher degree than changing the initialization.

3.3   MCC results of best solution

When using the "5-step heuristic" (run as part of stability analysis 1), would found that the final best MCC clustering model was for Run A (seed 360) which included four views as listed in Table 2. For visualizations, the subject and feature clusters were sorted according to their size and proportion of SZ patients as described in section 2.7. Figure 5 the views are visualized. View 1 and 2 (top) both included features that all were FC values, where the red color indicates positive connectivity, and blue indicates negative connectivity (i.e. anti correlation between time courses). View 3 and 4 only include binary features: view 3 included: gender, and the acquisition site of each participants while view 4: included the handedness, where blue lines are participants who had missing values. For the remaining sections we will not consider view 4, since this view only included a single feature and no subject clustering.
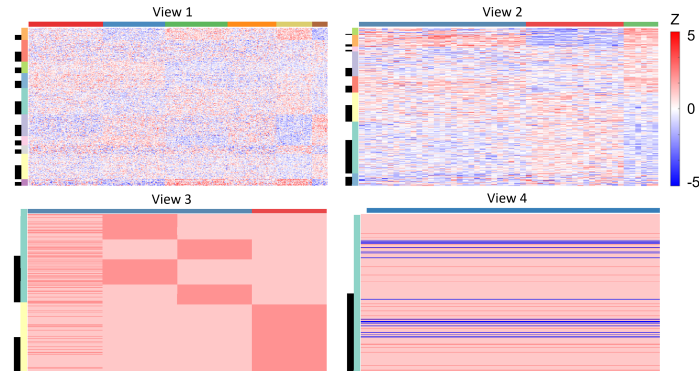
Fig. 5. Illustration of views from best model (Run A). For each view, the horizontal colorbar shows the feature clustering, while the vertical colorbar indicates the subject clusters. Patients with SZ are indicated with a black hyphen.)

3.4   Diagnosis association

To search for potential schizophrenia related subtypes, we used Pearson's χ^2 test for contingency tables to find views (and thereby subject clustering) with significant diagnosis association, as described in section 2.7. We found that view 2 had a significant diagnosis association with a high $\chi^2_{df=7} = 35$ (p < 0.001), whereas there was no significant diagnosis association for view 1 ($\chi^2_{df=9} = 1.8$ (p = 0.99)) nor for view 3 ($\chi^2_{df=1} = 0$ (p = 0.98)). This pattern was also clearly seen on Figure 5, where view 1 and view 3 had approximately the same number of patients with SZ (black hyphen) and healthy controls in each subject cluster, whereas most subject clusters in view 2 had a predominance of one of the groups (the proportion of SZ patients (SZ% is shown in Panel A in Figure 7). To gain a better understanding of the features included view 2, we show the binary FC matrices for each of the feature clusters in Figure 6. Feature cluster 3 contained six features, including connectivity between the basal ganglia RSN (gray) and the auditory (yellow), sensory motor (blue) , visual (purple) and default mode (red) resting state networks. On contrast, feature clusters 1 and 2 included more connectivity features which were also distributed between more RSN.

3.5   Separability between subject clusters

To gain further insight into the subject cluster differences within view 2, we used Cohen's D (CD) to asses the separability between neighboring clusters. Panel A in Figure 7 shows the proportion of SZ patients in each subject cluster, Panel B shows the histograms of each subject-feature cluster and Panel C shows the subject cluster separability measured by Cohen's D. We found a large mean (across feature clusters) CD between subject cluster (SC) 2-3 and between SC 5-6. The latter was particularly strong for feature cluster 3 (Cohen' D > 2). When we look at feature cluster 3, we see
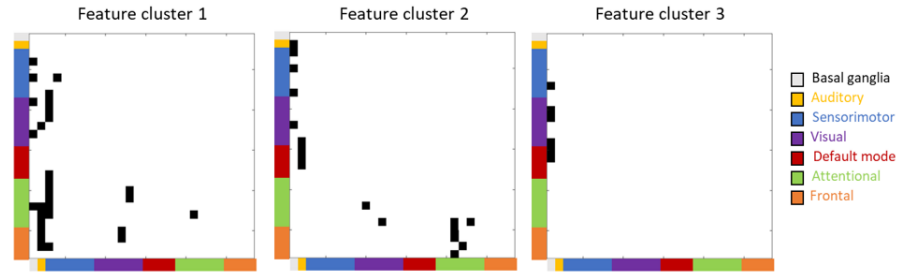
Fig. 6. FC matrices for the three feature cluster solutions of view 2. Binary FC matrices for each of the three feature clusters of View 2. The colorbars on the axes indicate the resting state network (RSN) assignment of each brain parcel according to the brain parcellation by Allen et al [27]

a linear trend with positive connectivity (red on Figure 5 and $\mu > 0$ for in Panel B of Figure 7) for subject clusters with a predominance of healthy controls (SC 1-5), while subject clusters which are dominated by SZ patients (SC 6-8) show negative connectivity valllues.

This is also seen in PDF of feature cluster 3 (Figure 8) where the subject cluster PDFs follows a linear trend from positive to negative according to the proportion of SZ in each cluster. For feature cluster 1, subject cluster 2 and 5 show a distinct pattern, while there was a large overlap of the remaining subject clusters.

3.6   Relation of clinical scales

Finally, we wanted to determine if the mean activation within each feature feature clusters were related to the clinical rating scales (PANSS) that we had available. We were particularly interested to see if the linear trend for feature cluster 3 would be related to the overall symptom severity measured by the total PANSS .

Figure 9 shows the scatter plots and correlation values (Rho) for each feature clusters, with the total PANSS (measure of symptom severity) and PANSS subscales (positive, negative and generalized). Each participant is colored according to their subject cluster assignment. Overall we see that the patients within a subject cluster are gathered around a horizontal line (indicating similar mean FC), which is in accordance with our expectation since subject clusters were formed based by participants with similar feature distributions. We did not find any feature clusters with a strong correlation between the mean feature activation and any of the four PANSS scales. And none of the correlations were significant after correction for multiple comparisons (using maximum permutation statistics over the three feature clusters and four PANSS scales). The strongest correlation was found for feature cluster 2 and the negative (neg) PANSS subscale ($\rho = 0.25$ p = 0.01 (uncorrected)). Furthermore, we found a weak positive trend (non adjusted p-value = 0.07, Rho = 0.16) between the mean connectivity values of feature cluster 3 (which included basal ganglia RSNs) and the positive PANSS subscale.
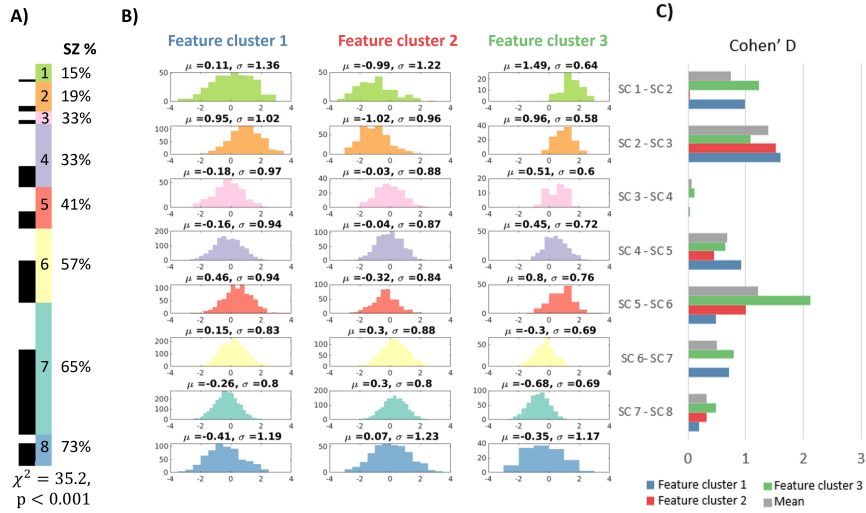
Fig. 7. Subject cluster separability for View 2. Panel A shows the proportion of patients with schizophrenia (SZ) in each of the subject clusters. Panel B shows the histogram of the FC values for each subject-feature cluster, and Panel C shows the separability between subject clusters measured by Cohens D.

Fig. 8. Probability density function (PDF) for each subject cluster of View 2.. The PDF of the subject clusters for each feature cluster. The PDF is determined by the hyper-parameters for each subject-feature cluster and the PDF given in Eq. 5.
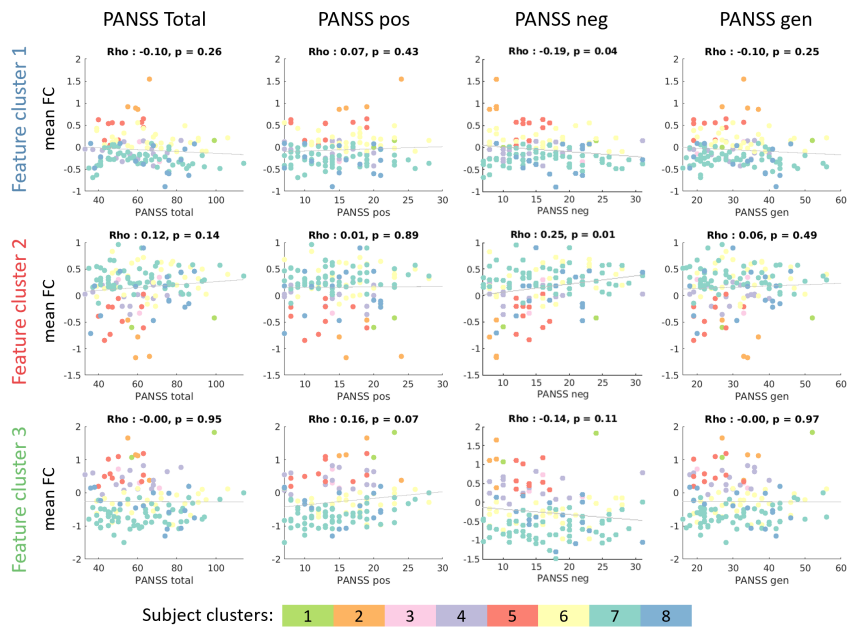
Fig. 9. Correlation between feature clusters activation and clinical scales. Scatter plots between the PANSS scale (PANSS total (measure of symptom severity), and three PANSS subscales (positive (pos), negative (neg) and generalized (gen)) and the mean value of the feature clusters. The correlation is measured using Spearman's rank coefficient (rho) and significance is assessed using a random permutation test (here not corrected for multiple comparisons). Each participant is colored according to their subject cluster assignment, and the gray line indicate a linear regression line.

With respect to the subject clusters, we found that nearly all patients with a PANSS total > 75 (considered "moderately affected by schizophrenia" [32]) were located in the last three subject clusters, indicating that the first five subject clusters (which also have a predominance of HC) do not include patients with a high symptom severity. However, we found that a large part of the patients from the last three subject clusters include patients with low symptom severity. A similar trend is found for the three PANSS subscales.

3.7 Reproducible on independent test data (D2)

To determine the reproducibility of the diagnosis association for view 2, we performed a validation analysis of the independent test dataset (D2) as described in section 2.8.
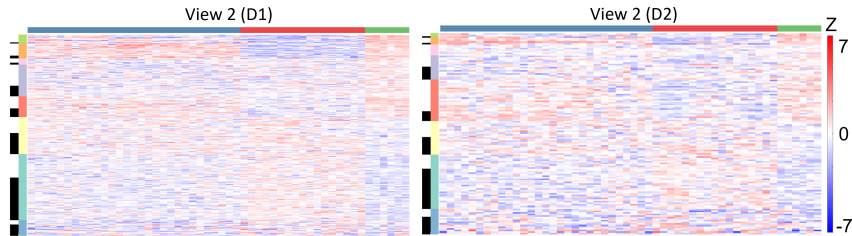
Fig. 10. View 2 for discovery (D1) and Test (D2) datasets. Visualization of View 2 (significant diagnosis association) for the two datasets. View 2 (D1) is the same as in Figure 5 but with an adjusted colormap matched the one in View (D2).

Figure 10 shows view 2 for the discovery (D1) and independent test (D2) datasets. Overall, it is seen that the views are similar across the datasets, but with higher connectivity strength for the D2 dataset. The main result of the reproducibility analysis was that the diagnosis association between the subject clusters and diagnostic label was still significant $\chi^2_{df=7} = 25.8$ (p < 0.001) on the independent test dataset (D2) (Panel A of Figure 11. Figure 11 further shows that a similar trend for the subject clusters are found, both in relation to proportion of SZ patients (Panel A) the histograms (Panel B) and subject cluster separability (Panel C). Furthermore, feature cluster 3 also still showed a red-blue linear trend with increasing subject clusters, whereas this was less clear for feature cluster 2.

Figure 12 shows the correlations between the feature clusters in the test view, and the four PANSS scores. The positive correlation between feature cluster 2 and the negative PANSS scale which was found on the discovery dataset, did not replicate on the external data. Similarly, the positive trend between feature cluster 3 and the positive PANSS scale was also not replicated. On the contrary, for the test dataset, feature cluster 3 showed a moderate negative correlation ($\rho \leq -0.30, p \leq 0.05$ (uncorrected) for all PANSS scales.
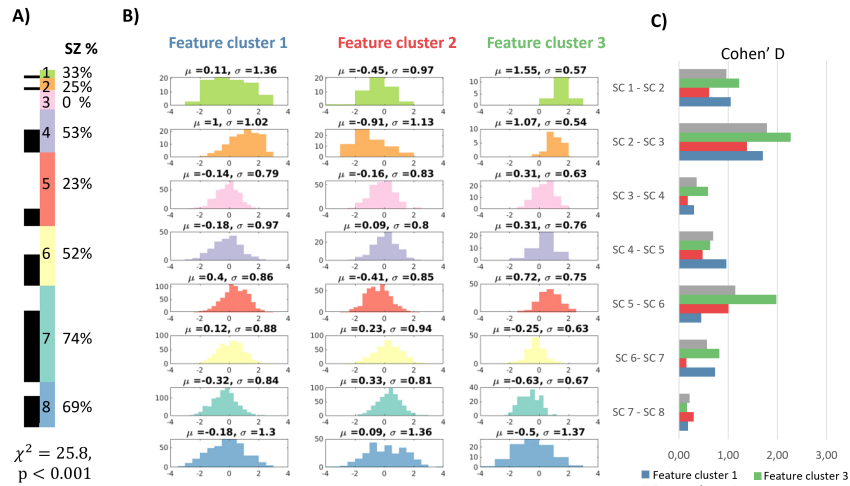
Fig. 11. Subject cluster separability for View 2. Panel A shows the proportion of patients with schizophrenia (SZ) in each of the subject clusters. Panel B shows the histogram of the FC values for each subject-feature cluster, and Panel C shows the separability between subject clusters measured by Cohen's D.

4   Discussion

4.1   Multiple co-clustering result

The final solution (Run A) of the multiple co-clustering (MCC) included 4 views, of which 2 views only included binary features representing handedness (view 4), as well as gender and data acquisition site (view 3) respectively. Since view 4 only included a single feature (handedness) and one subject cluster, no further analysis was performed on that view. View 3 included two subject and two feature clusters. Interpreting the included features, showed that this co-clustering structure related mostly to the site of the participants. I.e., feature cluster 1 includes the handedness (feature 1) and index for two of the sites (feature 2 and 3, both from the DecNef database), whereas the second feature cluster includes data from the remaining site (feature 4, COBRE database). This finding illustrates the advantage of the multiple views , which allows the MCC algorithm to determine several subject clustering solutions on the same dataset. For example, a clustering method which is only able to find one subject clustering solution (which needs to fit all the data), could have chosen a solution that relates to differences in acquisition site, gender or handedness instead of subtypes within a disease. On the contrary, the MCC method can put these into separate views. Furthermore, if there would have been brain connectivity features, which would be highly related to e.g. site differences, these could be clustered together with
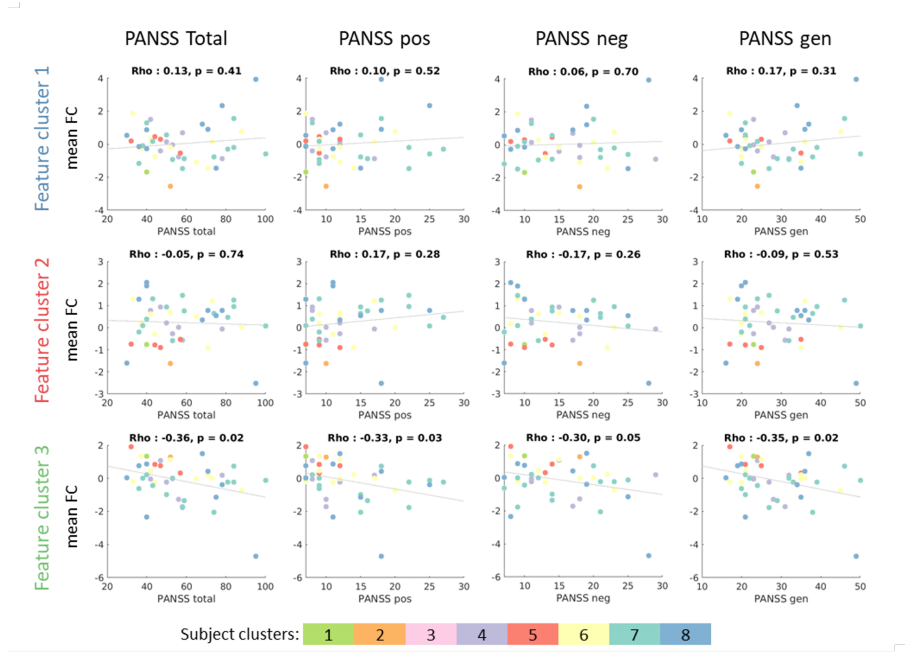
Fig. 12. Correlation between feature clusters activation and clinical scales for D2. Scatter plots between the PANSS scale (PANSS total (measure of symptom severity), and three PANSS subscales (postive (pos), negative (neg) and generalized (gen)) and the mean value of the feature clusters. The correlation is measured using Spearman's rank coefficient (rho) and significance is assessed using a random permutations test (not corrected for multiple comparisons). Each participant is colored according to their subject cluster assignment, and the gray line indicate a linear regression line.

the site-indicator features (polytopic learning), and thus separating these from the remaining connectivity features.

4.2   Interpretation of the stability analyses

In the stability analysis we determined the clustering similarity between different initializations (analysis 1) and datasplits (analysis 2). First of all, we found that the similarity was much lower for the runs with different datasplits (ARI_{view} = 0.48, Figure 3), than when re-running the clustering on the same dataset with different initialization (ARI_{view} = 0.84, Figure 4). This was in accordance with our anticipation, indicating that the clustering is more stable across initializations compared to the stability caused by variability on the data. If this would not have been the case, this would have raised severe concerns regarding the stability of the model estimation procedure.

For both stability analyses, we still found occasional abrupt changes in the log likelihood close to the maximum number of iterations (1000), indicating that the algorithm was not yet fully converged. Originally, we started with 30 iterations (which is the default setting in the implementation by Tokuda et al. [21]) but after inspecting the log-likelihood, we gradually increased the number of iterations until 1000. We did not include more iterations due to the high computational and time complexity. In earlier publications using the MCC algorithm, the number of iterations were not specified, and information about the cost function trajectory was not included, which makes it difficult to compare the convergence of our work with earlier results. Comparing the log-likelihood of the two stability analyses (Panel A in Figure 3 and 4 respectively) we find that the increases in the cost function over iterations are relatively small compared to the differences between data splits. However, since the datasets are not the same for these analyses, direct comparison of the cost functions is not possible. For future studies, we strongly recommend to investigate a broad range of iterations, and to report the number of iterations and cost functions when sharing the results.

To gain a better understanding of the subject clustering stability, the second part of the stability analysis 1 focused on the subject and feature clustering within the views on the top-two-pair runs. We found that these two runs had a very high stability on the view level (ARI_{view} = 0.99), and furthermore the feature clustering stability within each view was also high (ARI_{view} 0.85 and 0.96 for view 1 and view 2 as specified in Table 2). However, even though the feature clustering was stable between these runs, the subject clustering stability was somewhat lower with a ARI_{view} 0.77 and 0.67 for for view 1 and view 2 respectively. This finding indicates, that even for the most stable runs (step 4 of the 5-step heuristic presented in section 2.6) the subject clustering, and thereby potential subtype stability, was moderate.

To the best of our knowledge, this is the first study to report the clustering similarity of the MCC algorithm across initializations and datasplits. Earlier studies also investigated 1000 initialization of the algorithm and used the ARI to choose the "best" solution (as part of the 5-step heuristic); however, they did not report the actual stability values [14, 20]. The closest comparisons we have been able to find, were from a paper by Tokuda et al. from 2018, where they in the supplementary material report

the ARI across changes in the hyper-parameters (and concluded that these were stable, with ARI > 0.80). Furthermore, in Tokuda et al from 2017, they used the ARI on simulated data to determine the difference to the true solution.

In summary, our stability analysis showed that the stability of the MCC algorithm on the given datasets was higher between initializations that with datasplits. However, even though the feature to view clustering was nearly identical for the top-two pair of runs, the subject clustering similarity was only moderate. This indicates that the subject clusters, and thereby potential subtypes, are not very robust even across initializations.

4.3   Diagnosis association and evaluation of subtypes

In the remaining sections we focus on the potential subtypes that were found by the "best" solution (Run A) found in stability analysis 1. First we discuss the interpretation of the subject clustering solution found by view 2, and then we will further discuss the findings related to feature cluster 3.

4.4   Interpretation of subject clusters:

We found a reproducibly significant diagnosis association for the subject clusters in view 2 ($\chi^2_{df=7} = 35$ (p < 0.001) on the discovery data and ($\chi^2_{df=7} = 26$ (p < 0.001) on the external test data). This view included three feature clusters and eight subject clusters. Three of these subject clusters had a higher proportion of SZ patients (>50 %) and could thus be considered "SZ-related clusters" (subject cluster 6−8) compared to five "healthy control clusters" (subject cluster 1−5). However, it should be noted that there were still both SZ patients and HC in all clusters, and the diagnosis separation between these clusters were therefore not as clear as in the earlier subtyping work on patients with major depressive disorder, where a very clear separation was found between the disorder and control clusters [12].

Looking at the PANSS scales, we found that nearly all patients with a PANSS total > 75 (considered at least moderately affected by schizophrenia [32]) were located in these three "SZ-related clusters". This shows that even though the "healthy control clusters" included SZ patients, these patients had relatively low symptom severity.

To gain further insight into the separability between the clusters, we used the histograms, PDFs and effect sizes of clustering differences for each subject-feature cluster (Figure 7). Here, we found a high subject cluster separability between subject cluster 5 and 6, which was particularly strong for feature cluster 3 (Cohen's D > 2).

4.5   Interpretation of feature clusters 3

Overall view 2 included three feature clusters, as illustrated in Figure 6. Feature cluster 3 showed a linear trend from positive (red) connectivity for HC-clusters, and negative (blue) connectivity for SZ-related clusters. This was most clearly observed for the plot of the subject cluster PDFs (Figure 8) and could also be visually seen in Figure 10 and in the plot of the histograms (Figure 7). To determine if this linear trend was

related to the symptom severity of the SZ patients, we performed a correlation analysis as shown in Figure 9. We did not find any solid (significant after multiple comparison correction) correlation between any of the feature clusters and the PANSS scale, neither for the total PANSS (reflecting overall symptom severity) nor PANSS subscales. This shows that the feature clusters did not directly relate to any of the clinical representations that were available through the PANSS scale, which indicates that the MCC clustering reflects other sources of variability in the data.

As shown in Figure 6, all six connectivity features included in feature cluster 3, were related to the basal ganglia RSN. The basal ganglia RSN of the Allen atlas included several subcortical regions, hereby the striatum [27]. Many earlier studies have suggested that the striatum has a core role in schizophrenia, particularly in relation to the dopamine hypothesis [33–35] and positive symptoms [36]. In our correlation analysis, we found that there was a weak positive correlation on the discovery dataset (Figure 9 , Rho = 0.16, $p_{uncorrected}$ = 0.07) between the $PANSS_{total}$ and the mean activation of feature cluster 3. However, this was not reproduced on the external data. On the contrary, for the test dataset, feature cluster 3 showed a moderate negative correlation (Rho ≤ -0.30 $p_{uncorrected}$ ≤ 0.05) for all PANSS scales. Since the Allen atlas only included one RSN with subcortical areas, we can not draw any firm conclusion on the role of the striatum nor its correlations to the PANSS scales. However we see this as a promising finding which would be interesting to investigate in future studies, using a more fine grained atlas to study the influence of different subcortical regions.

To summarize, we found a moderately stable subject clustering solution which had a significant diagnosis association for view 2. Within this view, we found three "SZ-related" subject clusters with a predominance of SZ patients, and one feature cluster that showed a linear trend from positive to negative correlation with an increasing number of SZ patients in each subject cluster. None of the feature clusters showed a solid relation of the clinical manifestations measured through the PANSS scale. This indicates that the subject clusters may represent another differentiation than traditionally measured using the symptomatology based PANSS scale, however further exploration would be needed to draw firm conclusions. Importantly, the subject clustering and diagnosis association reproduced to the external test data, which indicates that the subject clustering was not only specific for the discovery dataset, but carries information that generalizes to data from independent sites.

4.6 Future directions

With regards to the stability of the clustering, we believe that even more structured investigation are needed. In future studies, we think it will be important to further investigate how the stability depends on various factors such as the availability of data (both the amount of participants included and features), different data types, number of iterations and other external factors (e.g., when including multi-site vs. single site data). For example, it would be interesting to see if the stability convergences with the number of iterations, and if this could be used as a cut off measure of the number of iterations, instead of using the default value or using convergence

of the log likelihood. We also suggest that future studies consider other stability metrics than the ARI. In the current work used the ARI to keep consistency with earlier publications. However, we note that the assumptions in permutation model can be violated [28, 37], since different runs can show different numbers of clusters (views, subject and feature clusters) and different numbers of features within each cluster. Hence, alternatives such as the permutation method presented by Gates et al. [37,38] could also be considered.

It would also be interesting to conduct a structured investigation of the posterior probabilities of the subject-feature clustering (given by the MCC clustering) and to compare these results to the stability analysis, to see if the clustering is more stable when the model has a high posterior probability (as a measure of certainty) for the subject-feature cluster assignments.

In this study, we used a relatively small RSN brain parcellation which included 28 RSN. In future applications it would be interesting to expand the analysis with a more fine grained atlas, e.g. to be able to specifically study the role of striatum. It should be noted that the stability analysis should be repeated on a new atlas, since an increase in the feature space dimension will likely influence the stability.

Furthermore, whereas we have investigated if there were any correlations between the clustering solutions and the PANSS scale, there are many other factors that could be interesting to include in a post-hoc analysis if available. For example, it would be interesting to see if any of the subject clusters were related to external factor such as smoking or treatment history, which we did not have available information about in our study.

In this study we included both patients with SZ and healthy controls, since this enabled us to evaluate views with a diagnosis association (as in earlier studies [12, 20]). However, with the goal of finding potential subtypes within patients with the heterogeneous SZ diagnosis, we see a interesting potential for running the MCC only on patients with SZ and then also including clinically relevant information, such as PANSS scores and treatment history if available, to take better advantage of the polytopic learning potential of the MCC algorithm.

5   Conclusion

The goal of this study was to use the multiple co-clustering method on functional connectivity data, to search for data-driven schizophrenia subtypes.

We found that the stability was higher across initializations than datasplits, but that even for models with a high view and feature clustering similarity, the subject clustering similarity was moderate. This highlights the importance of studying the stability of potential subtyping methods, and we believe that even more efforts should be devoted to figuring out how the stability depends on various factors, such as the size of the input data and convergence of the algorithm. Nevertheless, we found a subtyping solution which had a significant diagnosis association both on the discovery and the external dataset. This clustering solution included three subtypes with a potential relation to schizophrenia, which had a predominance of schizophrenia patients on both datasets and included nearly all patients with a moderate to

high PANSS score. Furthermore, we found a feature cluster where the connectivity values showed a linear trend on both datasets, and where all features were related to the basal ganglia RSN. Finally, none of the feature clusters were reliably correlated to any of the PANSS scales, which indicates that the clusters reflect other sources of variability in the data. In future studies, it would be very interesting to study if a similar subtype can be identified using a more fine-grained atlas, which would enable investigation of the potential influence of specific regions such as the striatum.

We see these results as very promising steps, and consider subtyping methods, such as the MCC method, to have great potential for exploring more data-driven disease subtypes for patients with schizophrenia.

References

1. L Miranda, R Paul, B Pütz, N Koutsouleris, and B Müller-Myhsok. Systematic review of functional mri applications for psychiatric disease subtyping. *Front Psychiatry*, 12:665536, 2021. 1664-0640 Miranda, Lucas Paul, Riya Pütz, Benno Koutsouleris, Nikolaos Müller-Myhsok, Bertram Systematic Review Switzerland Front Psychiatry. 2021 Oct 22;12:665536. doi: 10.3389/fpsyt.2021.665536. eCollection 2021.

2. Ana Vilar, Víctor Pérez-Sola, María Jesús Blasco, Elena Pérez-Gallo, Laura Ballester Coma, Santiago Batlle Vila, Jordi Alonso, Antoni Serrano-Blanco, and Carlos G Forero. Translational research in psychiatry: The research domain criteria project (rdoc). *Revista de Psiquiatría y Salud Mental (English Edition)*, 12:187–195, 2019.

3. T Insel, B Cuthbert, M Garvey, R Heinssen, D S Pine, K Quinn, C Sanislow, and P Wang. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders. *Am J Psychiatry*, 167:748–751, 2010. 1535-7228 Insel, Thomas Cuthbert, Bruce Garvey, Marjorie Heinssen, Robert Pine, Daniel S Quinn, Kevin Sanislow, Charles Wang, Philip Journal Article United States Am J Psychiatry. 2010 Jul;167(7):748-51. doi: 10.1176/appi.ajp.2010.09091379.

4. N V Kraguljac, W M McDonald, A S Widge, C I Rodriguez, M Tohen, and C B Nemeroff. Neuroimaging biomarkers in schizophrenia. *Am J Psychiatry*, 178:509–521, 2021. 1535-7228 Kraguljac, Nina V McDonald, William M Widge, Alik S Rodriguez, Carolyn I Tohen, Mauricio Nemeroff, Charles B K23 MH106683/MH/NIMH NIH HHS/United States R01 MH118484/MH/NIMH NIH HHS/United States R01 MH119384/MH/NIMH NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Review Am J Psychiatry. 2021 Jun;178(6):509-521. doi: 10.1176/appi.ajp.2020.20030340. Epub 2021 Jan 5.

5. W.D. Penny, K.J. Friston, J.T. Ashburner, S.J. Kiebel, and T.E. Nichols. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier Science, 2011.

6. Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols. *Handbook of Functional MRI Data Analysis*. Cambridge University Press, 2011.

7. V D Calhoun, G D Pearlson, and J Sui. Data-driven approaches to neuroimaging biomarkers for neurological and psychiatric disorders: emerging approaches and examples. *Curr Opin Neurol*, 34:469–479, 2021. 1473-6551 Calhoun, Vince D Pearlson, Godfrey D Sui, Jing R01 MH118695/MH/NIMH NIH HHS/United States R01 MH117107/MH/NIMH NIH HHS/United States R01 MH123610/MH/NIMH NIH HHS/United States UL1 TR001863/TR/NCATS NIH HHS/United States RF1 AG063153/AG/NIA NIH HHS/United States R01 EB020407/EB/NIBIB NIH HHS/United States R01 AG073949/AG/NIA NIH HHS/United States R01 EB006841/EB/NIBIB NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Review England Curr Opin Neurol. 2021 Aug 1;34(4):469-479. doi: 10.1097/WCO.0000000000000967.

8. M R Arbabshirani, S Plis, J Sui, and V D Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145:137–165, 2017. 1095-9572 Arbabshirani, Mohammad R Plis, Sergey Sui, Jing Calhoun, Vince D P20 GM103472/GM/NIGMS NIH HHS/United States R01 DA040487/DA/NIDA NIH HHS/United States R01 EB005846/EB/NIBIB NIH HHS/United States R01 EB006841/EB/NIBIB NIH HHS/United States R01 EB020407/EB/NIBIB NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Review United States Neuroimage. 2017 Jan 15;145(Pt B):137-165. doi: 10.1016/j.neuroimage.2016.02.079. Epub 2016 Mar 21.

9. Andrea Mechelli and Sandra Vieira. From models to tools: clinical translation of machine learning studies in psychosis. *npj Schizophrenia*, 6, 2020.

10. H. G. Schnack and R. S. Kahn. Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Front Psychiatry*, 7:50, 2016. 1664-0640 Schnack, Hugo G Kahn, René S Journal Article Switzerland Front Psychiatry. 2016 Mar 31;7:50. doi: 10.3389/fpsyt.2016.00050. eCollection 2016.

11. Darrel A. Regier, William E. Narrow, Diana E. Clarke, Helena C. Kraemer, S. Janet Kuramoto, Emily A. Kuhl, and David J. Kupfer. Dsm-5 field trials in the united states and canada, part ii: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, 170(1):59–70, 2013.

12. Tomoki Tokuda, Junichiro Yoshimoto, Yu Shimizu, Go Okada, Masahiro Takamura, Yasumasa Okamoto, Shigeto Yamawaki, and Kenji Doya. Identification of depression subtypes and relevant brain regions using a data-driven approach. *Scientific Reports*, 8, 2018.

13. Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, Alan F Schatzberg, Keith Sudheimer, Jennifer Keller, Helen S Mayberg, Faith M Gunning, George S Alexopoulos, Michael D Fox, Alvaro Pascual-Leone, Henning U Voss, Bj Casey, Marc J Dubin, and Conor Liston. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, 23:28–38, 2017.

14. Tomoki Tokuda, Okito Yamashita, and Junichiro Yoshimoto. Multiple clustering for identifying subject clusters and brain sub-networks using functional connectivity matrices without vectorization. *Neural Networks*, 142:269–287, 2021.

15. Grigorios F. Tzortzis and Aristidis C. Likas. Multiple view clustering using a weighted combination of exemplar-based mixture models. *IEEE Transactions on Neural Networks*, 21(12):1925–1938, 2010.

16. Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics Data Analysis*, 71:52–78, 2014.

17. Andre F Marquand, Thomas Wolfers, Maarten Mennes, Jan Buitelaar, and Christian F Beckmann. Beyond lumping and splitting: A review of computational approaches for stratifying psychiatric disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1:433–447, 2016.

18. Kay H Brodersen, Lorenz Deserno, Florian Schlagenhauf, Zhihao Lin, Will D Penny, Joachim M Buhmann, and Klaas E Stephan. Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage: Clinical*, 4:98–111, 2014.

19. Zhi Yang, Yong Xu, Ting Xu, Colin W Hoy, Daniel A Handwerker, Gang Chen, Georg Northoff, Xi-Nian Zuo, and Peter A Bandettini. Brain network informed subject community detection in early-onset schizophrenia. *Scientific Reports*, 4, 2015.

20. T. Tokuda, O. Yamashita, Y. Sakai, and J. Yoshimoto. Clustering of multiple psychiatric disorders using functional connectivity in the data-driven brain subnetwork. *Front Psychiatry*, 12:683280, 2021. 1664-0640 Tokuda, Tomoki Yamashita, Okito Sakai, Yuki Yoshimoto, Junichiro Journal Article Switzerland Front Psychiatry. 2021 Aug 18;12:683280. doi: 10.3389/fpsyt.2021.683280. eCollection 2021.

21. T. Tokuda, J. Yoshimoto, Y. Shimizu, G. Okada, M. Takamura, Y. Okamoto, S. Yamawaki, and K. Doya. Multiple co-clustering based on nonparametric mixture models with heterogeneous marginal distributions. *PLoS One*, 12(10):e0186566, 2017. 1932-6203 Tokuda, Tomoki Orcid: 0000-0001-6284-2113 Yoshimoto, Junichiro Shimizu, Yu Okada, Go Takamura, Masahiro Okamoto, Yasumasa Yamawaki, Shigeto Doya, Kenji Journal Article United States PLoS One. 2017 Oct 19;12(10):e0186566. doi: 10.1371/journal.pone.0186566. eCollection 2017.

22. Saori C Tanaka, Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, Naohiro Okada, Ryuichiro Hashimoto, Go Okada, Yuki Sakai, Jun Morimoto, Jin Naru-moto, Yasuhiro Shimada, Hiroaki Mano, Wako Yoshida, Ben Seymour, Takeshi Shimizu, Koichi Hosomi, Youichi Saitoh, Kiyoto Kasai, Nobumasa Kato, Hidehiko Takahashi, Ya-sumasa Okamoto, Okito Yamashita, Mitsuo Kawato, and Hiroshi Imamizu. A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data*, 8, 2021.

23. The Center for Biomedical Research Excellence (COBRE).

24. Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

25. Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth Dupre, Madeleine Sny-der, Hiroyuki Oya, Satrajit S Ghosh, Jessey Wright, Joke Durnez, Russell A Poldrack, and Krzysztof J Gorgolewski. fmriprep: a robust preprocessing pipeline for functional mri. *Nature Methods*, 16:111–116, 2019.

26. Karl J Friston, Steven Williams, Robert Howard, and Richard S J Frackowiak. Movement related effects in fmri. *Magnetic Resonance in Medicine*, 3:346–355, 1996.

27. Elena Allen, Erik Erhardt, Eswar Damaraju, William Gruner, Judith Segall, Rogers Silva, Martin Havlicek, Srinivas Rachakonda, Jill Fries, Ravi Kalyanam, Andrew Michael, Arvind Caprihan, Jessica Turner, Tom Eichele, Steven Adelsheim, Angela Bryan, Juan Bustillo, Vincent Clark, Sarah Feldstein Ewing, Francesca Filbey, Corey Ford, Kent Hutchison, Rex Jung, Kent Kiehl, Piyadasa Kodituwakku, Yuko Komesu, Andrew Mayer, Godfrey Pearlson, John Phillips, Joseph Sadek, Michael Stevens, Ursina Teuscher, Robert Thoma, and Vince Calhoun. A baseline for the multivariate comparison of resting-state networks. *Frontiers in Systems Neuroscience*, 5, 2011.

28. Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

29. Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15:1–25, 2002.

30. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):26–, 2009.

31. *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY, 2008.

32. Stefan Leucht, John M. Kane, Werner Kissling, Johannes Hamann, Eva Etschel, and Rolf R. Engel. What does the panss mean? *Schizophrenia Research*, 79(2):231–238, 2005.

33. Oliver D Howes and Shitij Kapur. The dopamine hypothesis of schizophrenia: version iii—the final common pathway. *Schizophrenia bulletin*, 35(3):549–562, 2009.

34. Jodi J Weinstein, Muhammad O Chohan, Mark Slifstein, Lawrence S Kegeles, Holly Moore, and Anissa Abi-Dargham. Pathway-specific dopamine abnormalities in schizophrenia. *Biological psychiatry*, 81(1):31–42, 2017.

35. D Jacobs and T Silverstone. Dextroamphetamine-induced arousal in human subjects as a model for mania. *Psychological medicine*, 16(2):323–329, 1986.

36. Robert A McCutcheon, Tiago Reis Marques, and Oliver D Howes. Schizophrenia—an overview. *JAMA psychiatry*, 77(2):201–210, 2020.
37. Alexander J. Gates and Yong-Yeol Ahn. The impact of random models on clustering similarity. *J. Mach. Learn. Res.*, 18(1):3049–3076, jan 2017.
38. Alexander J. Gates and Yong-Yeol Ahn. Clusim: a python package for calculating clustering similarity. *J. Open Source Softw.*, 4:1264, 2019.