



Deep sr-DDL: Deep structurally regularized dynamic dictionary learning to integrate multimodal and dynamic functional connectomics data for multidimensional clinical characterizations

N.S. D'Souza ^{a,*}, M.B. Nebel ^{b,c}, D. Crocetti ^b, J. Robinson ^b, N. Wymbs ^{b,c}, S.H. Mostofsky ^{b,c,d}, A. Venkataraman ^{a,*}

^a Department of Electrical and Computer Engineering, Johns Hopkins University, USA

^b Center for Neurodevelopmental & Imaging Research, Kennedy Krieger Institute, USA

^c Department of Neurology, Johns Hopkins School of Medicine, USA

^d Department of Psychiatry and Behavioral Science, Johns Hopkins School of Medicine, USA

ARTICLE INFO

Keywords:

Dynamic dictionary learning
Structural regularization
Multimodal integration
Functional magnetic resonance imaging
Diffusion tensor imaging
Clinical severity

ABSTRACT

We propose a novel integrated framework that jointly models complementary information from resting-state functional MRI (rs-fMRI) connectivity and diffusion tensor imaging (DTI) tractography to extract biomarkers of brain connectivity predictive of behavior. Our framework couples a generative model of the connectomics data with a deep network that predicts behavioral scores. The generative component is a structurally-regularized Dynamic Dictionary Learning (sr-DDL) model that decomposes the dynamic rs-fMRI correlation matrices into a collection of shared basis networks and time varying subject-specific loadings. We use the DTI tractography to regularize this matrix factorization and learn anatomically informed functional connectivity profiles. The deep component of our framework is an LSTM-ANN block, which uses the temporal evolution of the subject-specific sr-DDL loadings to predict multidimensional clinical characterizations. Our joint optimization strategy collectively estimates the basis networks, the subject-specific time-varying loadings, and the neural network weights. We validate our framework on a dataset of neurotypical individuals from the Human Connectome Project (HCP) database to map to cognition and on a separate multi-score prediction task on individuals diagnosed with Autism Spectrum Disorder (ASD) in a five-fold cross validation setting. Our hybrid model outperforms several state-of-the-art approaches at clinical outcome prediction and learns interpretable multimodal neural signatures of brain organization.

1. Introduction

Functional magnetic resonance imaging (fMRI) quantifies the changes in blood flow and oxygenation in the regions associated with neuronal activity. More specifically, resting state fMRI (rs-fMRI) is acquired in the absence of a task paradigm, thus allowing us to probe the spontaneous co-activation patterns in the brain. It is believed that the co-activations reflect the intrinsic functional connectivity between brain regions (Fox and Raichle, 2007). In contrast to fMRI, Diffusion Tensor Imaging (DTI) (Assaf and Pasternak, 2008) assesses structural connectivity by measuring the diffusion of water molecules across neuronal fibres in the brain. Going one step further, we can use tractography to construct detailed 3D maps of anatomical pathways within the brain based on the diffusion tensors. There is strong evidence in literature of the correspondence between functional and structural pathways within

the brain (Skudlarski et al., 2008), with several studies suggesting that this functional connectivity may be mediated by either direct or indirect anatomical connections (Atasoy et al., 2016; Bowman et al., 2012; Fukushima et al., 2018; Honey et al., 2009). Thus, rs-fMRI and DTI data provide complementary information about function and structure respectively, which when integrated together can be used to construct a more comprehensive view of brain organization both in health and disease. As a result, multimodal integration has become an important topic of study for the characterization of neuropsychiatric disorders such as Autism Spectrum Disorder (ASD) (Vissers et al., 2012), Attention Deficit Hyperactivity Disorder (ADHD) (Weyandt et al., 2013), and Schizophrenia (Niznikiewicz et al., 2003).

Traditional multimodal analyses of rs-fMRI and DTI data have largely focused on post-hoc statistical comparisons of features extracted from the data. For example, simple statistical differences in rs-fMRI and DTI connectivity between subjects have been used to discover disrupted pat-

* Corresponding author.

E-mail address: Shimona.Niharika.Dsouza@jhu.edu (N.S. D'Souza).

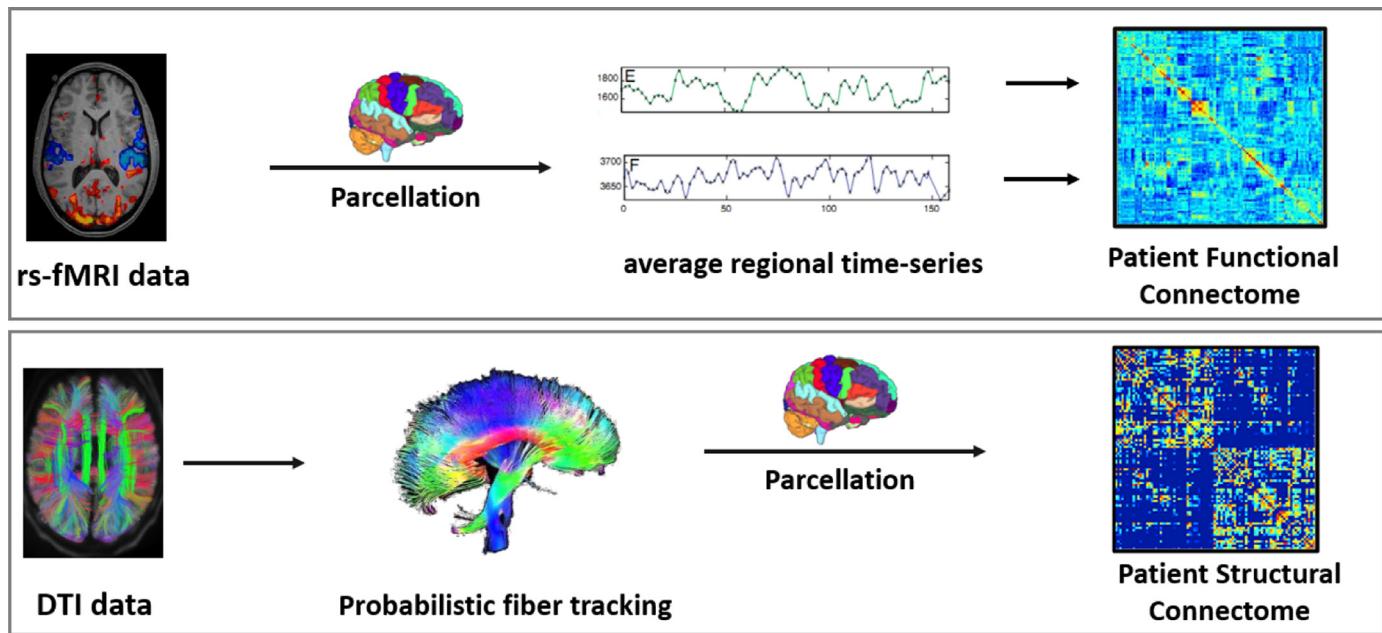


Fig. 1. Top: For the fMRI data, we group voxels in the brain into ROIs defined by a standard atlas and compute the average time courses for each ROI. The correlation matrix captures the synchrony in the average time courses. **Bottom** Tractography is performed on the raw DWI data to track the path of neuronal fibers in the brain. Based on the parcellation scheme, we construct a map of the fibre tracts between ROIs in the brain. The same parcellation scheme is used for both modalities.

terns of brain organization in Alzheimer's disease (Hahn et al., 2013) and Progressive Supranuclear Palsy (PSP) (Whitwell et al., 2011). On a population level, classical multivariate analysis (Andrews-Hanna et al., 2007; Goble et al., 2012) or random effects models (Propper et al., 2010) are employed to independently compute and then combine features from both modalities. Despite their past success at biomarker discovery, these techniques often fail to generalize at a patient-specific level. Furthermore, they often ignore higher-order interactions between multiple subsystems in the brain, which is known to be critical for understanding complex neuropsychiatric disorders (Kaiser et al., 2010; Koshino et al., 2005). These shortcomings have paved the way for the development of the network based view of brain connectivity that simultaneously accounts for both inter-subject and intra-subject variability.

In the case of fMRI, network-based models often group voxels in the brain into regions of interest (ROIs) using a standard anatomical or functional atlas. Next, the functional relationships between these regions are determined based on the synchrony between representative (often average) regional time series. This information is typically represented in terms of a static functional connectivity matrix as shown in Fig. 1 (top). In case of DTI, tractography is used to estimate the fiber tracts between the ROIs in the brain from the voxel-level diffusion tensors, from which features such as the anisotropy or the number of fibers can be extracted. Similar to the functional connectome, the structural connectivity matrix captures the strength of the pairwise anatomical connection between different ROIs, as seen in Fig. 1 (bottom).

Some of the simplest approaches to analyzing network properties borrow heavily from the field of graph theory. For example, the works of (Bullmore and Sporns, 2009; Rubinov and Sporns, 2010; Sporns et al., 2004) use aggregate network measures, such as node degree, betweenness centrality, and eigenvector centrality to study the organization of the brain. These measures compactly summarize the connectivity information onto a restricted set of nodes that can be mapped back to the brain. A more global network property is small-worldness (Bassett and Bullmore, 2006), which describes an architecture of sparsely connected clusters of nodes. Complementary changes in small-worldness in both anatomical and functional networks have been well documented across the literature (Park et al., 2008; Sun et al., 2014), with concurrent disruptions of functional networks (Wang et al., 2009) or structural net-

works (Wang et al., 2012) implicated in neuropsychiatric disorders such as schizophrenia. The main limitation of these approaches is that they independently analyze the fMRI and DTI data, and as such, draw heuristic conclusions about the relationship between the two modalities.

Community detection techniques have been widely used for understanding the organization of complex systems such as the brain (Bardella et al., 2016). Other examples include the work of Venkataraman et al. (2013) that identifies abnormal connectivity in schizophrenia, and (Venkataraman et al., 2016), which characterizes the social and communicative deficits associated with autism. An alternative network topology is the hub-spoke model, used by Venkataraman et al. (2013), Venkataraman et al. (2012) and Venkataraman et al. (2015), that targets regions associated with a large number of altered rs-fMRI connections. These methods, however, exclusively focus on functional connectivity and do not incorporate structure. In this light, the work of Venkataraman et al. (2011) proposes a probabilistic framework that jointly models latent anatomical and functional connectivity to discover population-level differences in schizophrenia. Similarly, the work of Higgins et al. (2018) uses a unified Bayesian framework to identify gender-differences in multimodal connectivity patterns across different age groups. While successful at combining multi-modal information for group differentiation, these techniques do not directly address inter-individual variability.

Data-driven methods integrating structural and functional connectivity focus heavily on groupwise discrimination from the static connectomes. These methods usually follow a two-step approach where feature selectors and discriminators are trained sequentially in a pipeline. For example, the authors in Wee et al. (2012) combine graph theoretic features computed from rs-fMRI and DTI graphs with Support Vector Machines (SVMs) to identify individuals with Mild Cognitive Impairment. Another example is the work of Sui et al. (2013), which employs a pipeline consisting of joint-Independent Component Analysis (j-ICA) on the two modalities followed by Canonical Correlation Analysis (CCA) to combine them and distinguish schizophrenia patients from controls. In contrast to the pipelined approaches, end-to-end deep learning methods combining feature selection and prediction are becoming ubiquitous in neuroimaging studies. These are highly successful due to their ability to learn complex abstractions directly from input data. As an example,

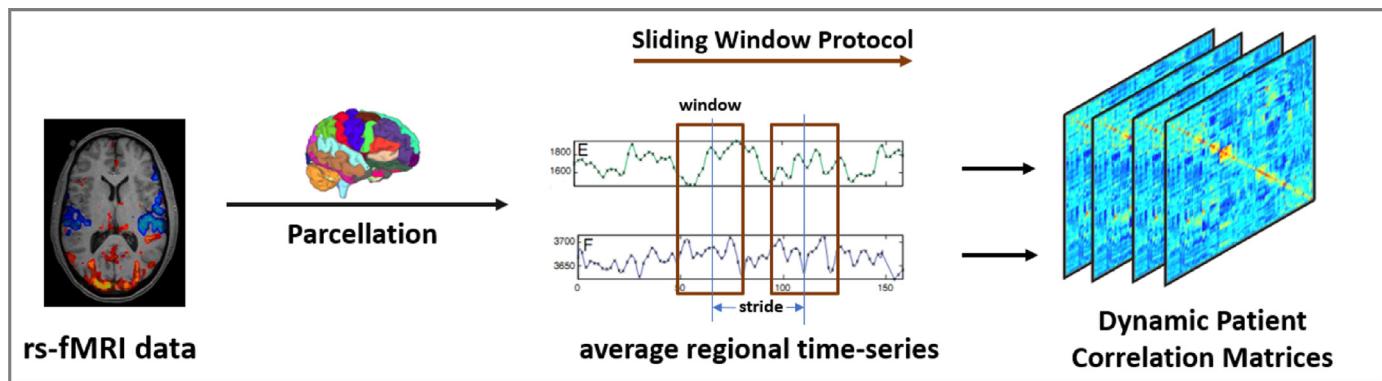


Fig. 2. First, the ROI's defined by a standard atlas are used to compute regional time series. Then, a sliding window protocol defined by window length and stride is applied to extract the dynamic patient correlation matrices. As in the static case, the dynamic matrices measure the synchrony between regional time series, but as a function of time.

the work of [Aghdam et al. \(2018\)](#) uses a Deep Belief Network (DBN) on multimodal data to disambiguate patients with Autism Spectrum Disorder from healthy controls. However, none of the above methods tackle continuous-valued prediction, for example, quantifying a continuous level of deficit.

In the continuous prediction realm, our previous works in [D'Souza et al. \(2018\)](#) and [D'Souza et al. \(2019a\)](#) combine dictionary learning on rs-fMRI correlation matrices with linear and non-linear regression models respectively to predict a single measure of clinical severity. These methods combine the rs-fMRI representation with the prediction in a coupled optimization framework. Unfortunately, they fail to generalize to predicting multiple deficits (i.e. multi-score prediction). While they use a similar coupled optimization strategy, they fail to generalize to predicting multiple deficits (i.e. multi-score prediction). On the other hand, recent works of [Kawahara et al. \(2017\)](#) and [D'Souza et al. \(2019b\)](#) have demonstrated the power of deep neural networks to map to multiple clinical/cognitive outcomes from rs-fMRI and DTI data separately. While promising, all of these methods focus on a single neuroimaging modality and do not exploit complementary interactions between structural and functional connectivity. In addition, the aforementioned techniques rely on static rs-fMRI correlation matrices as input. Consequently, they largely ignore the dynamics of evolution of the functional scan.

There is now growing evidence that functional connectivity is a dynamic process that toggles between different intrinsic states evolving over a static structural connectome ([Cabral et al., 2017](#)). These states manifest over short time windows that are typically of the order of a tens of seconds to a few minutes. Several studies such as [Rashid et al. \(2014\)](#) and [Price et al. \(2014\)](#) indicate the importance of modeling this evolution for characterizing neuropsychiatric disorders such as schizophrenia and Autism Spectrum Disorder (ASD). The dynamic connectivity among ROIs in the brain is typically captured via a sliding window protocol, defined by the window length and stride, as illustrated in [Fig. 2](#). The window length defines the length of the time sequence considered by each dynamic correlation matrix, while the stride controls the overlap in successive sliding windows. Recently, model based alternatives that detect dynamic changes in correlation between large-scale brain networks such as the Default Mode Network, Somatosensory Network etc have been developed. An example is the Dynamic Conditional Correlation (DCC) protocol that was initially developed in the econometrics and finance literature ([Engle, 2002](#)) and later adapted to the study of brain organization using rs-fMRI ([Lindquist, 2016](#)). It poses a time-varying matrix estimation problem to explicitly model the evolution of connectivity patterns in the brain, and has shown robustness in the test-retest setting ([Lindquist et al., 2014](#)) with rs-fMRI. Unfortunately, this method is unstable when scaled up ([Aielli, 2013](#); [Caporin and McAleer, 2013](#)), for example to a whole brain

ROI-level analysis of dynamic connectivity, likely due to ill conditioning of the correlation matrices in the absence of additional regularization. Consequently, most dynamic connectivity studies continue to rely on sliding-window correlations as inputs. Examples include ([Cai et al., 2017](#)), where the authors use a sparse decomposition of the rs-fMRI connectomes, or [Rabany et al. \(2019\)](#), which employs a temporal clustering for ASD/control discrimination. Nevertheless, these approaches focus exclusively on rs-fMRI and completely ignore structural information.

We propose a deep-generative hybrid model, i.e. the deep sr-DDL, that integrates structural and dynamic functional connectivity with behavior into a unified optimization framework.

1.1. Our contribution

The contributions of this work are two-fold. From an application standpoint, we develop a unified framework to integrate structural (DTI) and dynamic rs-fMRI connectivity together with behavior. From a technical standpoint, we propose a unique alternative to black-box deep learning methods by combining the interpretability of classical techniques with the representational power of strategically-designed deep neural networks. As a starting point, we leverage the dictionary learning frameworks of [Eavani et al. \(2015\)](#), [D'Souza et al. \(2018\)](#) and [D'Souza et al. \(2019a,b\)](#), which extract group-level subnetworks from static rs-fMRI correlation matrices. Our deep sr-DDL carries this method further via two main components:

- A generative dictionary learning component to represent the multi-modal and dynamic data
- A deep network to model the temporal trends and predict behavioral scores.

Our generative component is a structurally regularized Dynamic Dictionary Learning (sr-DDL), which uses a DTI tractography prior to regularize a matrix factorization of the dynamic rs-fMRI correlation matrices. The sr-DDL decomposes dynamic rs-fMRI correlation matrices into a collection of shared bases, and time-varying subject specific loadings. These loadings are input to a deep network which is comprised of a Long-Short Term Memory (LSTM) module to model temporal trends and an ANN that predicts clinical scores. The key to this generative-deep hybrid is our coupled optimization procedure, which jointly estimates the bases, loadings, and neural network weights most predictive of the individual behavioral profile.

A preliminary version of our work was published in MICCAI 2020 ([D'Souza et al., 2020b](#)). In this journal, we provide a detailed analysis of our framework where we validate on both synthetic data and two separate real-world datasets. The first of these includes a subset of healthy adults from the publicly available Human Connectome Project (HCP) ([Van Essen et al., 2012](#)). This helps us evaluate the efficacy of our

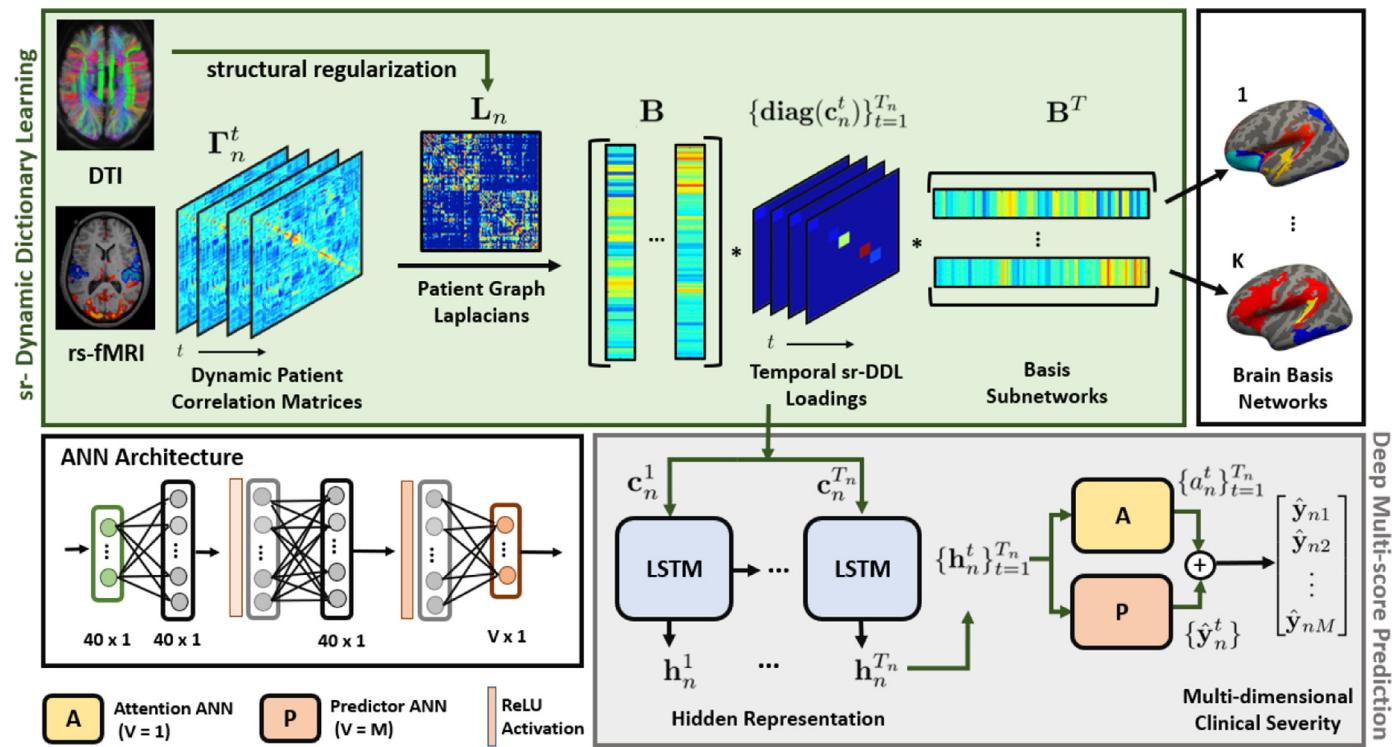


Fig. 3. Framework to integrate structural and dynamic functional connectivity for clinical severity prediction. **Green Box:** The generative sr-DDL module. The rs-fMRI dynamic correlation matrices are decomposed into the subnetwork basis and time-varying subject-specific loadings. The DTI connectivity regularizes this decomposition. **Purple Box:** Deep LSTM-ANN module for multi-score prediction. The sr-DDL coefficients are input into the LSTM to generate a hidden representation. The predictor ANN (P-ANN) generates a time varying estimate for the scores, while the attention ANN (A-ANN) weights the predictions across time to generate the final clinical severity estimate.

framework at predicting cognitive outcomes from the rs-fMRI and DTI scans. Next, we examine a clinical dataset consisting of children diagnosed with Autism Spectrum Disorder (ASD). The presentation of ASD is known to be heterogeneous with individuals exhibiting a wide spectrum of behavioral impairments in terms of social reciprocity, communicative functioning, and repetitive/restrictive behaviours (Spitzer and Williams, 1980), quantified via clinical severity measures. We observed that our method outperforms several state-of-the-art approaches at predicting behavioral performance in unseen individuals from their connectomics data for both datasets. This illustrates that our method is reproducible. Furthermore, we provide a detailed presentation of our clinical results, especially the subnetworks identified by the model in both datasets. We conclude with a discussion on the generalizability, and robustness and potential directions of future work.

In summary, our joint objective balances generalizability with interpretability, bridging the representational gap between structure, function and behavior. Our experiments highlight the potential of our deep sr-DDL framework for providing a more holistic view of neuropsychiatric diseases.

2. Materials and methods

2.1. A deep generative hybrid model to integrate multimodal and dynamic connectivity with behavior

Fig. 3 presents a graphical overview of our framework. We have two sets of inputs to the model for each individual namely, the dynamic individual-specific correlation matrices, and the DTI structural connectome graph (upper left). Our outputs are the scalar clinical scores (bottom right). We use the sliding window approach in Fig. 2 to extract dynamic rs-fMRI correlation matrices and tractography to extract the DTI connectomes as shown in Fig. 1. The DTI input to our model is the

Graph Laplacian obtained from a binary DTI adjacency matrix capturing the presence/absence of a fiber between regions. Finally, the behavioral scores for each individual are obtained from an expert assessment. This score can correspond to either cognitive outcomes or severity of symptoms in case of neurodevelopmental diseases.

The green box in Fig. 3 describes the generative component of our framework. Here, the dynamic rs-fMRI correlation matrices are decomposed using a structurally regularized dynamic dictionary learning (sr-DDL). The columns in the bases subnetworks capture representative patterns common to the cohort. The loading coefficients differ across subjects, and evolve over time. At each timepoint/observation, they determine the contribution of each basis to the dynamic functional connectivity profile of the individual. Finally, the DTI Graph Laplacians re-weight the decomposition to focus on the functional connectivity between anatomically linked regions. The gray box denotes the deep networks part of our model. This network combines a Long Short Term Memory (LSTM) module with an Artificial Neural Network (ANN) to predict multiple behavioral scores. The LSTM models the temporal trends in the subject-specific loading coefficients giving rise to a hidden representation. The ANN then uses this representation to predict the corresponding behavioral outcomes.

Dynamic Dictionary Learning for rs-fMRI data:

We denote the set of time varying functional correlation matrices for individual n by the set $\{\Gamma_n^t\}_{t=1}^{T_n} \in \mathcal{R}^{P \times P}$. Here, T_n denotes the number of sliding windows applied to the rs-fMRI scan, and P is the number of ROIs in the parcellation scheme. As seen in Fig. 3 (green box), we model this information using a group average basis, and subject-specific temporal loadings. The dictionary $\mathbf{B} \in \mathcal{R}^{P \times K}$ is a concatenation of K elemental bases vectors $\mathbf{b}_k \in \mathcal{R}^{P \times 1}$, i.e. $\mathbf{B} := [\mathbf{b}_1 \ \mathbf{b}_2 \ \dots \ \mathbf{b}_K]$, where $K \ll P$. This basis captures representative brain states which each subject cycles through over the course of the scan. We further constrain the basis vectors to be orthogonal to each other. This constraint acts as an implicit

regularizer, ensuring that the learned subnetworks are uncorrelated, yet explain the rs-fMRI data well. While the bases are shared across the cohort, the strength of their combination differs across individuals and varies over time. These loadings are denoted by the set $\{\mathbf{c}_n^t\}_{t=1}^{T_n}$ and combine the basis subnetworks uniquely to best explain each subject's functional connectivity. We introduce an explicit non-negativity constraint \mathbf{c}_{nk}^t to ensure that the positive semi-definiteness of $\mathbf{\Gamma}_n^t$ is preserved. The complete rs-fMRI data representation takes the following form:

$$\mathbf{\Gamma}_n^t \approx \sum_k \mathbf{c}_{nk}^t \mathbf{b}_k \mathbf{b}_k^T \quad \text{s.t.} \quad \mathbf{c}_{nk}^t \geq 0, \quad \mathbf{B}^T \mathbf{B} = \mathbf{I}_K, \quad (1)$$

where \mathbf{I}_K is the $K \times K$ identity matrix. As seen in Eq. (1), the subject-specific loading vector at time t , $\mathbf{c}_n^t := [\mathbf{c}_{n1}^t \dots \mathbf{c}_{nK}^t]^T \in \mathcal{R}^{K \times 1}$ models the heterogeneity in the cohort. Denoting $\text{diag}(\mathbf{c}_n^t)$ as a diagonal matrix with the K subject-specific coefficients on the diagonal and off-diagonal terms set to zero, Eq. (1) can be re-written in the following matrix form:

$$\mathbf{\Gamma}_n^t \approx \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T \quad \text{s.t.} \quad \mathbf{c}_{nk}^t \geq 0, \quad \mathbf{B}^T \mathbf{B} = \mathbf{I}_K \quad (2)$$

Finally, this matrix factorization serves to reduce the dimensionality of the rs-fMRI data, while simultaneously modeling group-level and subject-specific information.

Structural Regularization from DTI data: Let $\mathbf{A}_n \in \mathcal{R}^{P \times P}$ be a binary adjacency matrix derived from the structural connectome of subject n . For example, \mathbf{A}_n can be constructed by thresholding the number of fibers estimated between two regions via tractography. Let \mathcal{E} denote the set of edges in this graph. We compute the corresponding Normalized Graph Laplacian [Banerjee and Jost (2008)] as $\mathbf{L}_n = \mathbf{V}_n^{-\frac{1}{2}} (\mathbf{V}_n - \mathbf{A}_n) \mathbf{V}_n^{-\frac{1}{2}}$, where $\mathbf{V}_n = \text{diag}(\mathbf{A}_n \mathbf{1})$ is the degree matrix and $\mathbf{1}$ is the vector of all ones. Intuitively, the Graph Laplacian is a discrete analog of the Laplace difference operator in Euclidean space. The Laplace difference operator has been used to characterize local properties of functions in Euclidean space (for example, to easily identify and characterize local optima). The Graph Laplacian generalizes this notion to discrete graphs and functions that are defined on graphs. Specifically, the Graph Laplacian has become a popular spatial regularizer in computer vision (Pang and Cheung, 2017), genetics (Feng et al., 2017) and neuroimaging (Atasoy et al., 2016; Cuingnet et al., 2012). This regularization implicitly assumes that there is a data signal associated with each node of the graph, and it encourages these signals to be similar for nodes of the graph that have an edge between them.

We use a matrix analog to Graph Laplacian regularization via the weighted Frobenius norm i.e. $\|\cdot\|_{\mathbf{L}_n}$ (Manton et al., 2003; Schnabel and Toint, 1983), which we use in place of the isotropic ℓ_2 penalty in Eq. (2). In this case, the graph "signal" corresponds to the vector (i.e., profile) of approximation errors given in Eq. (2) between the node in question and all other nodes in the graph. The underlying anatomical connectivity graph is defined by the DTI Graph Laplacian \mathbf{L}_n for each patient. Mathematically, our dictionary learning loss takes the following form:

$$\begin{aligned} \|\mathbf{\Gamma}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T\|_{\mathbf{L}_n} \\ = \text{Tr}[(\mathbf{\Gamma}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T) \mathbf{L}_n (\mathbf{\Gamma}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T)] \end{aligned} \quad (3)$$

Here, $\text{Tr}[\mathbf{M}]$ is the trace operator, which sums the diagonal elements of the argument matrix \mathbf{M} . For convenience, let $\mathbf{E}_n^t = \mathbf{\Gamma}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T$ denote the element-wise approximation error of the correlation matrix $\mathbf{\Gamma}_n^t$. Similarly, we define $\tilde{\mathbf{E}}_n^t = \mathbf{V}_n^{-\frac{1}{2}} \mathbf{E}_n^t$ as a weighted version of this error based on the degree matrix. As detailed in Appendix A, Eq. (3) can be expanded as follows:

$$\begin{aligned} \|\mathbf{\Gamma}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T\|_{\mathbf{L}_n} \\ = \sum_{(i,k) \in \mathcal{E}} \|\tilde{\mathbf{E}}_n^t(i, :) - \tilde{\mathbf{E}}_n^t(k, :)\|_2^2 \\ = \sum_{(i,k) \in \mathcal{E}} \|[\mathbf{V}_n(i, i)]^{-\frac{1}{2}} \mathbf{E}_n^t(i, :) - [\mathbf{V}_n(k, k)]^{-\frac{1}{2}} \mathbf{E}_n^t(k, :)\|_2^2 \end{aligned} \quad (4)$$

Notice that for terms where $(i, k) \notin \mathcal{E}$, i.e. there is no anatomical connection between nodes i and k , the corresponding error term in the summation drops out. Said another way, this construction minimizes the sum of the square difference between the rs-fMRI reconstruction profiles ($\tilde{\mathbf{E}}_n^t(i, :)$ and $\tilde{\mathbf{E}}_n^t(k, :)$) between nodes $(i$ and k) that are adjacent via the DTI graph. This effectively re-weights the rs-fMRI reconstruction profiles of anatomically connected nodes according to their relative degrees ($\mathbf{V}_n(i, i)$ and $\mathbf{V}_n(k, k)$) in the DTI graph pairwise. Thus, the functional connectivity at a particular node is directly influenced by its anatomical connections with other nodes in the graph. At a high level, this construction implicitly regularizes the rs-fMRI reconstruction loss according to the underlying anatomical connectivity prior.

Finally, based on the formulation in Eq. (3), the final sr-DDL objective $\mathcal{D}(\cdot)$ can be expressed as follows:

$$\begin{aligned} \mathcal{D}(\mathbf{B}, \{\mathbf{c}_n^t\}; \{\mathbf{\Gamma}_n^t\}, \mathbf{L}_n) \\ = \sum_t \frac{1}{T_n} \|\mathbf{\Gamma}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T\|_{\mathbf{L}_n} \\ \text{s.t.} \quad \mathbf{c}_{nk}^t \geq 0, \quad \mathbf{B}^T \mathbf{B} = \mathbf{I}_K \end{aligned} \quad (5)$$

Deep Multiscore Prediction: As seen in the gray box in Fig. 3, the subject-specific coefficients $\{\mathbf{c}_n^t\}$ are input to an LSTM-ANN to predict the clinical scores, as parametrized by the weights Θ . The M clinical scores for each individual are concatenated into a vector $\mathbf{y}_n := [\mathbf{y}_{n1} \dots \mathbf{y}_{nM}^T] \in \mathcal{R}^{M \times 1}$. The LSTM models the temporal variations in the coefficients $\{\mathbf{c}_n^t\}$ to generate a hidden representation $\{\mathbf{h}_n^t\}_{t=1}^{T_n}$. From here, the Predictor ANN (P-ANN) generates a time varying estimates of the scores $\{\hat{\mathbf{y}}_n^t\}_{t=1}^{T_n} \in \mathcal{R}^{M \times 1}$. At the same time, the Attention ANN (A-ANN) generates T_n scalars from the hidden representation. These are then softmax across time to obtain the attention weights: $\{a_n^t\}_{t=1}^{T_n}$. The final prediction is an attention-weighted average across the time estimates, which takes the following form:

$$\hat{\mathbf{y}}_n = \sum_t \hat{\mathbf{y}}_n^t a_n^t \quad (6)$$

Effectively, the attention weights determine which time points for each subject are most relevant for behavioral prediction. Additionally, they allow us to handle rs-fMRI scans of varying durations. Mathematically, we compute the multi-score prediction error $\mathcal{L}(\cdot)$ using the Mean Squared Error (MSE) loss function as follows:

$$\mathcal{L}(\{\mathbf{c}_n^t\}, \mathbf{y}_n; \Theta) = \|\hat{\mathbf{y}}_n - \mathbf{y}_n\|_F^2 = \left\| \sum_{t=1}^{T_n} \hat{\mathbf{y}}_n^t a_n^t - \mathbf{y}_n \right\|_F^2 \quad (7)$$

At a high level, the deep network distills the temporal information to best predict each subject's clinical profile.

We would like to highlight that our choice of the LSTM over a Recurrent Neural Network (RNN) allows us to track the temporal evolution of connectivity over longer horizons, while avoiding issues with convergence (Chung et al., 2014). Our two branched ANN in conjunction with the LSTM directly pools together time-varying estimates of clinical severity by focusing on the portions of the rs-fMRI scan most relevant to prediction. We notice that this construction naturally allows us to handle scans of varying length, while at same time obviating the need for additional sequence padding as would be required by a competing 1D CNN.

In Section 2.2, we will develop a coupled optimization procedure to jointly estimate our unknowns $\{\mathbf{B}, \{\mathbf{c}_n^t\}, \Theta\}$. We will show that our estimation procedure for the coefficients and neural network weights only relies on backpropagated gradients from the neural network loss and the parametric gradients from the dictionary learning. From the joint objective in Eq. (8), we can see that the choice of neural network architecture does not directly affect the dictionary learning gradients. So long as we can backpropagate the deep network loss to the coefficients \mathbf{c}_n^t , we can effectively adopt our optimization strategy to handle an alternative architecture. Said another way, our coupled optimization procedure is agnostic to the specific neural network choice.

Architectural Details: Our proposed ANN architecture is highlighted in the white box to the bottom left of Fig. 3. Our modeling choices carefully control for representational capacity and convergence of our coupled optimization procedure. Since the input to the network, i.e. the coefficient vector \mathbf{c}_n^t is essentially low dimensional, we opt for a two layered LSTM with the hidden layer width as 40. Both the P-ANN and the A-ANN are fully connected neural networks with two hidden layers of width 40. Since the A-ANN outputs a scalar, the width of its output layer is one, while that of the P-ANN is of size M , i.e. the number of behavioral scores. We use a Rectified Linear Unit (ReLU) as the activation function for each hidden layer, as we found that this choice is robust to issues with vanishing gradients and saturation that commonly confound the training of deep neural networks (Glorot et al., 2011).

Joint Objective for Multimodal Integration: We combine the complementary viewpoints in Eqs. (5) and (7) into a single joint objective below:

$$\begin{aligned} \mathcal{J}(\mathbf{B}, \{\mathbf{c}_n^t\}, \Theta; \{\Gamma_n^t\}, \mathbf{L}_n, \{\mathbf{y}_n\}) &= \underbrace{\sum_n \mathcal{D}(\mathbf{B}, \{\mathbf{c}_n^t\}; \{\Gamma_n^t\}, \mathbf{L}_n)}_{\text{sr-DDL loss}} + \lambda \underbrace{\sum_n \mathcal{L}(\Theta, \{\mathbf{c}_n^t\}; \mathbf{y}_n)}_{\text{deep network loss}} \\ &= \sum_n \sum_t \frac{1}{T_n} \|\Gamma_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T\|_{\mathbf{L}_n} \\ &\quad + \lambda \sum_n \mathcal{L}(\Theta, \{\mathbf{c}_n^t\}; \mathbf{y}_n) \quad \text{s.t. } \mathbf{c}_{nk}^t \geq 0, \quad \mathbf{B}^T \mathbf{B} = \mathbf{I}_K \quad (8) \end{aligned}$$

Here, λ is a hyperparameter than balances the tradeoff between the representation loss $\mathcal{D}(\cdot)$ and the prediction loss $\mathcal{L}(\cdot)$. $\{\mathbf{B}, \{\mathbf{c}_n^t\}, \Theta\}$ are the variables to optimize.

2.2. Coupled optimization strategy

We employ the alternating minimization technique in order to infer the set of hidden variables $\{\mathbf{B}, \{\mathbf{c}_n^t\}, \Theta\}$. Namely, we optimize Eq. (8) for each output variable, while holding the other unknowns constant.

We utilize the fact that there is a closed-form Procrustes solution for quadratic objectives of the form $\|\mathbf{M} - \mathbf{B}\|_F^2$ (Eversen, 1998). However, Eq. (8) is bi-quadratic in \mathbf{B} , so it cannot be directly applied. Therefore, we adopt the strategy in D'Souza et al. (2020a, 2019a, 2019b) of introducing $\sum_n T_n$ constraints of the form $\mathbf{D}_n^t = \mathbf{B} \text{diag}(\mathbf{c}_n^t)$. These constraints are enforced via the Augmented Lagrangian algorithm with corresponding constraint variables $\{\Lambda_n^t\}$. Thus, our objective from Eq. (8) now becomes:

$$\begin{aligned} \mathcal{J}_c &= \sum_{n,t} \frac{1}{T_n} \|\Gamma_n^t - \mathbf{D}_n^t \mathbf{B}^T\|_{\mathbf{L}_n} + \lambda \sum_n \mathcal{L}(\Theta, \{\mathbf{c}_n^t\}; \mathbf{y}_n) \\ &\quad + \sum_{n,t} \frac{\gamma}{T_n} \left[\text{Tr}[(\Lambda_n^t)^T (\mathbf{D}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t))] \right] \\ &\quad + \sum_{n,t} \frac{\gamma}{T_n} \left[\frac{1}{2} \|\mathbf{D}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t)\|_F^2 \right] \\ &\quad \text{s.t. } \mathbf{c}_{nk}^t \geq 0, \quad \mathbf{B}^T \mathbf{B} = \mathbf{I}_K \quad (9) \end{aligned}$$

The Frobenius norm terms $\|\mathbf{D}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t)\|_F^2$ regularize the trace constraints during the optimization. Observe that Eq. (9) is convex in the set $\{\mathbf{D}_n^t\}$, which allows us to optimize this variable via standard procedures. The constraint parameter is fixed at $\gamma = 20$, based on the guidelines in the literature (Nocedal and Wright, 2006).

Fig. 4 depicts our alternating minimization strategy. We describe each individual block in detail below. We refer the interested reader to Appendix B, which systematically delineates the supporting calculations from this section.

Step 1: Closed form solution for \mathbf{B} : Notice that Eq. (9) reduces to the following quadratic form in \mathbf{B} :

$$\mathbf{B}^* = \arg \min_{\mathbf{B}: \mathbf{B}^T \mathbf{B} = \mathbf{I}_K} \|\mathbf{M} - \mathbf{B}\|_F^2 \quad (10)$$

Given the singular value decomposition $\mathbf{M} = \mathbf{USV}^T$, we have the following closed form solution :

$$\mathbf{B}^* = \mathbf{UV}^T$$

where \mathbf{M} is computed as follows:

$$\begin{aligned} \mathbf{M} &= \sum_n \frac{1}{T_n} \sum_t (\Gamma_n^t \mathbf{L}_n + \mathbf{L}_n \Gamma_n^t) \mathbf{D}_n^t + \\ &\quad \sum_n \frac{1}{T_n} \left[\sum_t \frac{\gamma}{2} \mathbf{D}_n^t \text{diag}(\mathbf{c}_n^t) + \gamma \Lambda_n^t \text{diag}(\mathbf{c}_n^t) \right] \quad (11) \end{aligned}$$

Essentially, \mathbf{B} spans the anatomically weighted space of subject-specific dynamic correlation matrices.

Step 2: Updating the sr-DDL loadings $\{\mathbf{c}_n^t\}$: The objective \mathcal{J}_c in Eq. (9) decouples across subjects. Additionally, we can also incorporate the non-negativity constraint $\mathbf{c}_{nk}^t \geq 0$ by passing an intermediate vector $\tilde{\mathbf{c}}_n^t$ through a ReLU. The ReLU pre-filtering allows us to optimize an unconstrained version of Eq. (9), which can be done via the stochastic ADAM algorithm (Kingma and Ba, 2015). In essence, this optimization couples the parametric gradient from the augmented Lagrangians with the backpropagated gradient from the deep network (defined by fixed Θ). After convergence, the thresholded loadings $\mathbf{c}_n^t = \text{ReLU}(\tilde{\mathbf{c}}_n^t)$ are used in subsequent steps.

Step 3: Updating the Deep Network weights- Θ : We backpropagate the loss $\mathcal{L}(\cdot)$ to solve for the unknowns Θ . Notice that by dropping the contributions of the unknown value of \mathbf{y}_{nm} to the network loss during backpropagation using the ADAM (Kingma and Ba, 2015) algorithm, we can handle missing clinical data as well.

Step 4: Updating the Constraint Variables $\{\mathbf{D}_n^t, \Lambda_n^t\}$: We perform parallel primal-dual updates for the constraint pairs $\{\mathbf{D}_n^t, \Lambda_n^t\}$. Here, we cycle through the closed form update for \mathbf{D}_n^t and gradient ascent for Λ_n^t until convergence.

Step 5: Prediction on Unseen Data: In our cross-validated setting, we need to compute the sr-DDL loadings $\{\tilde{\mathbf{c}}^t\}_{t=1}^T$ for a new patient based on the training \mathbf{B}^* . Since we do not know the score $\bar{\mathbf{y}}$ for this patient, we remove the contribution $\mathcal{L}(\cdot)$ from Eq. (8) and assume the constraints $\bar{\mathbf{D}}^t = \mathbf{B}^* \text{diag}(\tilde{\mathbf{c}}^t)$ hold with equality, thus removing the Lagrangian terms. Essentially, the optimization for $\{\tilde{\mathbf{c}}^t\}$ reduces to decoupled quadratic programming (QP) objectives Q_t across time:

$$\begin{aligned} \tilde{\mathbf{c}}^{t+1} &= \arg \min_{\tilde{\mathbf{c}}^t} \frac{1}{2} (\tilde{\mathbf{c}}^t)^T \bar{\mathbf{H}} \tilde{\mathbf{c}}^t + \bar{\mathbf{f}}^T \tilde{\mathbf{c}}^t \quad \text{s.t. } \bar{\mathbf{A}} \tilde{\mathbf{c}}^t \leq \bar{\mathbf{b}} \\ \bar{\mathbf{H}} &= 2(\mathbf{B}^{*T} \bar{\mathbf{L}} \mathbf{B}^*); \\ \bar{\mathbf{f}} &= -[\mathcal{I}_K \circ (\mathbf{B}^{*T} (\bar{\mathbf{F}}^t \bar{\mathbf{L}} + \bar{\mathbf{L}} \bar{\mathbf{F}}^t) \mathbf{B}^*)] \mathbf{1}; \\ \bar{\mathbf{A}} &= -\mathcal{I}_K \bar{\mathbf{b}} = \mathbf{0} \end{aligned}$$

Where, \circ denotes the Hadamard product. Finally, we estimate $\bar{\mathbf{y}}$ via a forward pass through the LSTM-ANN.

Overall, our alternating minimization training procedure explicitly couples the Dictionary Learning (sr-DDL) and Deep Network (LSTM-ANN) blocks within the optimization. In contrast, the setup at test time consists of two steps, namely the coefficient update followed by a forward pass through the LSTM-ANN. We will demonstrate via our experiments (i.e. Section 3.2) that the coupled training is key to generalization. Finally, we discuss the effect of this difference between the training and testing procedures further in Section 4.1

2.2.1. Implementation details

Parameter Settings: In order to fix the hyperparameters for our model and the baselines, we make use of a second subset of 130 individuals from the HCP database (hereby referred to as HCP-2). Note that these individuals have no overlap with those used to characterize the performance in Section 3.2 to avoid biasing the results. First, we set aside 30 of these patients as a validation set to determine appropriate learning rates for our method and baselines. Recall that our deep-generative hybrid has two free parameters: namely the penalty λ , which controls the tradeoff between data representation and clinical prediction, and K ,

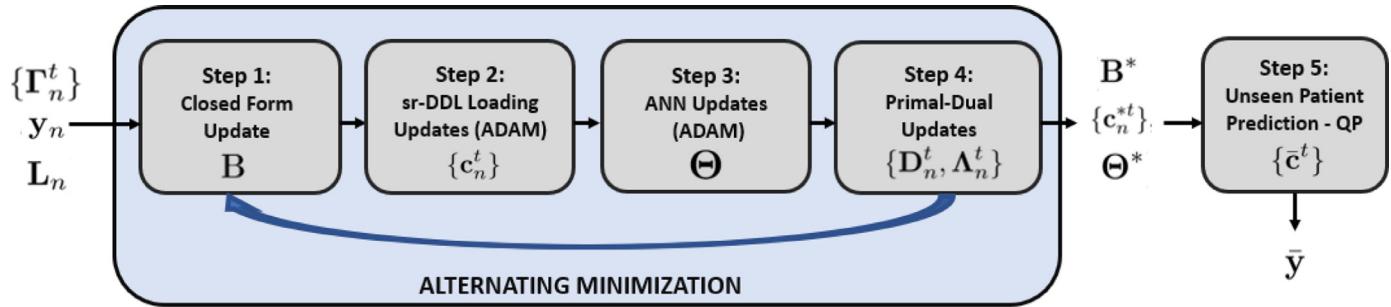


Fig. 4. Alternating minimization strategy for joint optimization of Eq. (9).

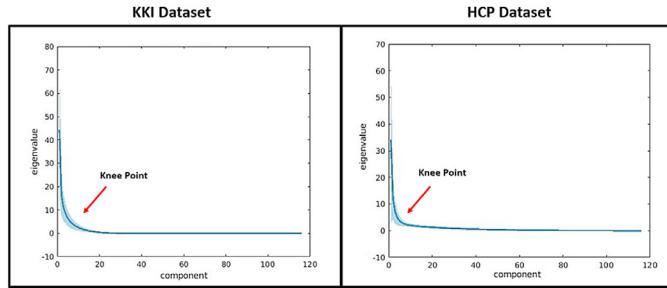


Fig. 5. Scree Plot of the correlation matrices to corroborate the selected values for K . (L) KKI Dataset (R) HCP Dataset. The thick line denotes the mean eigenvalue, while the shaded area indicates the standard deviation across subjects and time points.

the number of networks. For our experiments, we chose $K = 15$ for both datasets based on the knee point of the eigenspectrum of the correlation matrices $\{\Gamma_n^t\}$ (See Fig. 5). Based on the results of a 5 fold cross validation and grid search on HCP-2, we fix $\lambda = 2.5$. We will further discuss the robustness to λ in Section 4.2. Along similar lines, our Section 3.5 includes a discussion on emerging subnetwork patterns in \mathbf{B} upon varying the model order, i.e. K .

Additionally, our sliding window protocol is defined by two parameters, namely the window length and stride. Although these are not hyperparameters for the sr-DDL per se, they affect the predictive performance by controlling the information overlap between successive dynamic rs-fMRI correlation matrices. Again, these are set based on the cross validation performance on HCP-2. We will further discuss the robustness to these parameters in Section 4.2.

Initialization: Our coupled optimization strategy requires us to initialize the basis \mathbf{B} , coefficients $\{c_n^t\}$, the deep network weights Θ and the constraint variable pairs $\{D_n^t, \Lambda_n^t\}$. We randomly initialize the deep network weights at the first main iteration. We employ a soft-initialization for $\{\mathbf{B}, \{c_n^t\}\}$ by solving the dictionary objective in Eq. (5) without the LSTM-ANN loss terms for 20 iterations. We then initialize $\mathbf{D}_n^t = \mathbf{B} \text{diag}(\mathbf{c}_n^t)$ and $\Lambda_n^t = \mathbf{0}$ which lie in the feasible set for our constraints. We empirically observed that this soft initialization helps stabilize the optimization to provide improved predictive performance in fewer main iterations when compared with a completely random initialization.

Finally, the meta-data and code used in this study are available on a public repository hosted on Github¹.

2.3. Baseline comparison techniques

We evaluate the performance of our framework against three different classes of baselines, each highlighting the benefit of specific modeling choices made by our method.

Our first baseline class is a two stage configuration as illustrated in Fig. 6 that combines feature extraction on the dynamic rs-fMRI and DTI data, with a deep learning predictor. These feature engineering techniques are drawn from a set of well established statistical (Independent Component Analysis in Section 2.3.2) and graph theoretic techniques (Betweenness Centrality in Section 2.3.1), known to provide rich feature representations. The learned features are then input to the same deep LSTM-ANN network used by our method. This network is trained separately to predict the clinical outcomes. Note that these baselines incorporate multimodal and dynamic information, but do not directly operate on the network structure of the connectomes. Our second baseline class omits the two step approach in lieu of an end-to-end convolutional neural network based on the work of Kawahara et al. (2017). We train this model on the static rs-fMRI and DTI connectomes in tandem to predict the clinical scores. This baseline operates directly on the correlation and connectivity matrices, but ignores the dynamic evolution of functional connectivity. Next, we present the comparison of our deep sr-DDL by omitting the structural regularization. This helps us evaluate the benefit provided by the multimodal integration of DTI and rs-fMRI data. Our final baseline highlights the benefit of our joint optimization procedure. In this experiment, we decouple the optimization of the dynamic matrix factorization and deep network in Fig. 3 similar to the two stage pipelines.

2.3.1. Graph theoretic feature selection

Notice that the subject-specific correlation rs-fMRI matrices $\{\Gamma_n^t\}$ and the corresponding binary DTI adjacency matrices \mathbf{A}_n indicate time-varying functional and anatomical connectivity between the ROIs respectively. Therefore, we multiply the two to generate the time-varying multimodal graphs whose nodes are the brain ROIs and edges are defined by the temporal connectivity between these ROIs. We denote the corresponding adjacency matrices for these graphs by $\{\Psi_n^t = \mathbf{A}_n \circ \Gamma_n^t \in \mathcal{R}^{P \times P}\}$, where we threshold each Ψ_n^t to remove negative values. Each element $[\Psi_n^t]_{ij}$ gives the strength of association between two communicating sub-regions i and j in individual n at time t . We summarize the topology of these graphs via **Betweenness Centrality** (C_B) to obtain a time-varying estimate of brain connectivity for each ROI [Bassett and Bullmore (2006); Sporns et al. (2004)]. $C_B(v)$ for region v is calculated as:

$$C_B^t(v) = \sum_{s \neq v \neq u \in V} \frac{\sigma_{su}^t(v)}{\sigma_{su}^t} \quad (12)$$

σ_{su}^t is the total number of shortest paths from node s to node u at time t , and $\sigma_{su}^t(v)$ is the number of those paths that pass through v . This measure quantifies the number of times a node acts as a bridge along the shortest path between two other nodes and has found wide usage in characterizing small-worlded networks in brain connectivity (Sporns et al., 2004). We effectively reduce the dimensionality of the connectivity features. Again, the collection of features $\{C_B^t\}$ are used to train an LSTM-ANN predictor from Fig. 3 with two hidden layers having width 200 due to the higher input feature dimensionality.

¹ <https://github.com/Niharika-SD/Deep-sr-DDL>

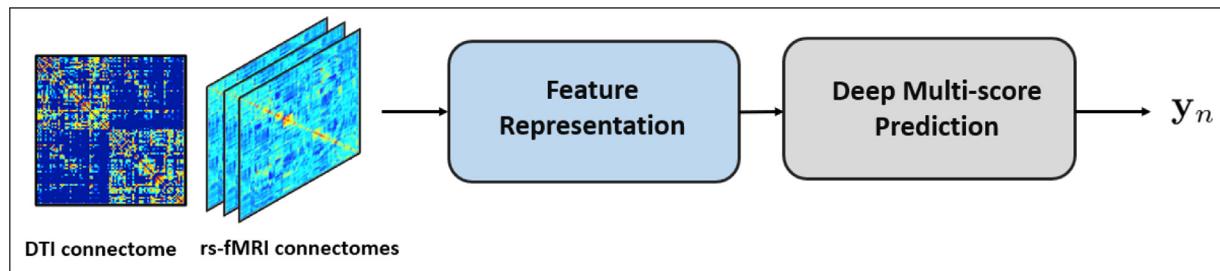


Fig. 6. A typical two stage baseline. We input the dynamic correlation matrices and DTI connectomes to Stage 1, which performs Feature Extraction. This step could be a technique from machine learning, graph theory or a statistical measure. Stage 2 is a deep network that predicts the clinical scores.

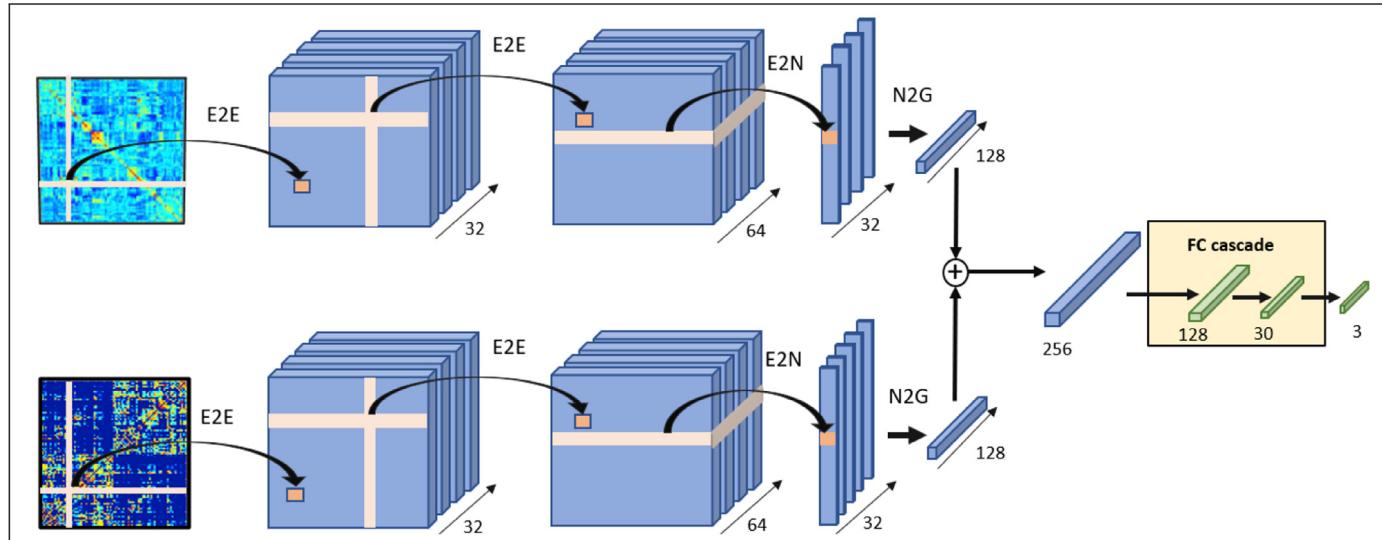


Fig. 7. The BrainNet CNN baseline (Kawahara et al., 2017) for severity prediction from multimodal data.

2.3.2. ICA feature selection

This baseline employs **Independent Component Analysis (ICA)** combined with the LSTM-ANN predictor. ICA is a statistical technique that extracts representative spatial patterns from the rs-fMRI time series. It has now become ubiquitous in fMRI analysis for its ability to identify group level differences as well as model individual-specific connectivity signatures. Essentially, ICA decomposes multivariate signals into ‘independent’ non-Gaussian components based on the data statistics.

This algorithm can be extended to the multi-subject analysis setting via Group ICA (G-ICA). Specifically, we extract independent spatial patterns common across patients, by combining the contribution of the individual time courses. For this baseline, we first perform G-ICA using the GIFT toolbox (Calhoun et al., 2009), and derive independent spatial maps for each subject from their raw rs-fMRI scans. We then compute the average time courses for each spatial map considering the constituent voxels. This provides us with a feature representation of reduced dimension equal to the number of specified maps ($d \ll L$) for each individual. For our experiments, we extract 15 ICA components. These time courses are input into the LSTM-ANN network in Fig. 3 with two hidden layers of width 40 to predict the clinical outcomes.

2.3.3. BrainNet convolutional neural network

The BrainNet CNN (Kawahara et al., 2017) relies on specialized fully convolutional layers for feature extraction, and was originally used to predict cognitive and motor outcomes from DTI connectomes. Fig. 7 provides a pictorial overview of the original architecture adapted for clinical outcome prediction from multimodal data. Each branch of the network accepts as input a $P \times P$ connectome, to which it applies a cascade of two edge-edge (E-E) convolutional operations. This E-E opera-

tion combines individual convolutions acting on the row and column to which the input element belongs. It is followed by a series of edge-node (E-N) blocks that reduce the dimensionality of the intermediate outputs, followed by a node-graph (N-G) operation for pooling. Finally, the output clinical scores are predicted via a fully connected artificial neural network for regression.

We feed the rs-fMRI static connectomes ($\hat{\Gamma}_n$) and DTI Laplacians \mathbf{L}_n into two disjoint fully convolutional branches with the architecture described above. We integrate the learned features via concatenation and input them into the fully connected layers described in Fig. 7, but with the number of outputs equal to the dimensionality of the clinical severity vector y_n . We set the learning rate, momentum and weight decay parameters according to the guidelines in Kawahara et al. (2017).

2.3.4. Deep sr-DDL without DTI regularization

In this baseline, we examine the effect of excluding the structural regularization provided by the DTI data from the joint objective in Eq. (8). The resulting objective function takes the following form:

$$\begin{aligned} \mathcal{J}_w(\mathbf{B}, \{\mathbf{c}_n^t\}, \Theta; \{\hat{\Gamma}_n^t\}, \{y_n\}) &= \sum_n \sum_t \frac{1}{T_n} \|\hat{\Gamma}_n^t - \mathbf{B} \text{diag}(\mathbf{c}_n^t) \mathbf{B}^T\|_F^2 \\ &\quad + \lambda \sum_n \mathcal{L}(\Theta, \{\mathbf{c}_n^t\}; y_n) \quad \text{s.t.} \quad \mathbf{c}_{nk}^t \geq 0, \quad \mathbf{B}^T \mathbf{B} = \mathbf{I}_K. \end{aligned} \quad (13)$$

Notice that amounts to replacing the Weighted Frobenius Norm formulation by a regular ℓ_2 penalty. This allows us to adopt the alternating minimization procedure in Section 2.2 to optimize Eq. (13) with a few minor modifications. Specifically, instead of T_n constraints per subject, we use a single constraint of the form $\mathbf{D} = \mathbf{B}$, enforced via a single Aug-

mented Lagrangian Λ . This effectively ensures that the new objective has a quadratic form in \mathbf{B} , along with a closed form update for \mathbf{D} . As before, we cycle through four individual steps, namely:

- Closed form Procrustes solution for the basis \mathbf{B}
- Updating the temporal loadings $\{\mathbf{c}_n^t\}$ (ADAM)
- Updating the Neural Network Parameters Θ (ADAM)
- Augmented Lagrangian updates for the constraint variables $\{\mathbf{D}, \Lambda\}$

Similar to the Deep sr-DDL, we use $K = 15$ networks as inputs to the LSTM-ANN network with two hidden layers of width 40 to predict the clinical outcomes.

2.3.5. Decoupled deep sr-DDL

Our final baseline examines the efficacy of our coupled optimization procedure in Section 2.2 with regards to generalization onto unseen subjects. Here, we first run the feature extraction using the sr-DDL optimization to extract the basis \mathbf{B} and temporal loadings $\{\mathbf{c}_n^t\}$. We then use the $\{\mathbf{c}_n^t\}$ as inputs to train the LSTM-ANN network in Fig. 3 to predict the scores \mathbf{y}_n . This is akin to the two-stage baselines delineated in Fig. 6.

Again, we use $K = 15$ networks with an a two layered LSTM-ANN having hidden layer width 40

3. Experimental results

3.1. Validation on synthetic data

As a sanity check, we first validate our optimization in Section 2.2 on synthetic data generated from the equivalent generative process. This experiment allows us to assess the behavior of our algorithm under various noise scenarios. Specifically, we evaluate the robustness of our estimation procedure under varying levels of noise in the correlation matrices and the scores, and under increasing deviations from orthogonality in our generating basis. Our simulations indicate that the optimization procedure is robust in the noise regime (0.01 – 0.2) estimated from the real-world rs-fMRI data. In addition, these experiments help us identify the stable parameter settings ($\lambda = 1 – 10$) which guide our real world experiments. We refer the interested reader to the Supplementary Results for the details from this section.

3.2. Real-world experiments: population studies of connectomics and behavior

We evaluate our deep-generative hybrid on two separate cohorts. The first dataset is a cohort of 150 healthy individuals from the Human Connectome Project (HCP) database (Van Essen et al., 2013) having both the rs-fMRI and DTI scans. We refer to this as the HCP dataset. Cognitive outcomes such as fluid intelligence are believed to be closely connected to structural (SC) and function connectivity (FC) in the human brain (Zimmermann et al., 2018). Thus, jointly modeling multi-modal neuroimaging and cognitive data helps exploit this fundamental interweave and uncover the neural underpinnings of cognition. Finally, we chose to focus on a modest sized dataset ($N = 150$) to demonstrate that our framework is suitable for clinical rs-fMRI applications, many of which have limited sample sizes.

Our second dataset consists of 57 children with high functioning Autism Spectrum Disorder (ASD) acquired at the Kennedy Krieger Institute in Baltimore, USA. Henceforth, we refer to this as the KKI dataset. The age of the subjects from this cohort is 10.06 ± 1.26 with an IQ of 110 ± 14.03 . Social and communicative deficits in ASD are believed to arise from aberrant interactions between regions of the brain that are linked by structural and functional connectivity (Rudie et al., 2013). Thus, identifying these patterns plays a crucial role in illuminating the etiological basis of the disorder.

Neuroimaging Data: As described in Van Essen et al. (2013), the HCP S1200 dataset was acquired on a Siemens 3T scanner (TR/TE = $0.72\text{ms}/0.33\text{ms}$, spatial resolution = $2 \times 2 \times 2\text{mm}$). The rs-fMRI scans

were processed according to the standard pre-processing pipeline described in Smith et al. (2013), which includes additional processing to account for confounds due to motion and physiological noise. We opted to use a 15 min interval (typical of clinical rs-fMRI studies of neurodevelopmental disorders) from the second scan of each subject's first visit for our analysis.

The DTI data from the HCP dataset was processed using the standard Neurodata MR Graphs package (ndmg) (Kiar et al., 2016). This consists of co-registration to anatomical space via FSL (Jenkinson et al., 2012), followed by tensor estimation in the MNI space and probabilistic tractography to compute the fibre tracking streamlines.

For the KKI dataset, rs-fMRI acquisition was performed on a Phillips 3T Achieva scanner with a single shot, partially parallel gradient-recalled EPI sequence with TR/TE = $2500/30\text{ms}$, flip angle 70° , res = $3.05 \times 3.15 \times 3\text{mm}$, having 128 or 156 time samples. The children were instructed to relax with eyes open and focus on a central cross-hair while remaining still. We used an in-house pre-processing pipeline pre-validated across several studies (D'Souza et al., 2020a; Nebel et al., 2016; Venkataraman et al., 2017). This consists of slice time correction, rigid body realignment, and normalization to the EPI version of the MNI template using SPM (Penny et al., 2011), followed by temporal detrending of the time courses to remove gradual trends in the data. A CompCorr50 (Ciric et al., 2018; Muschelli et al., 2014) strategy was used to estimate and remove spatially coherent noise from the white matter and ventricles, along with the linearly detrended versions of the six rigid body realignment parameters and their first derivatives, followed by spatial smoothing using a 6mm FWHM Gaussian kernel and temporal smoothing via a band pass filter (0.01 – 0.1Hz). Lastly, the data was despiked using the AFNI package (Cox, 1996).

The DTI acquisition for the KKI dataset was collected on a 3T Philips scanner (EPI, SENSE factor = 2.5, TR = 6.356s , TE = 75ms , res = $0.8 \times 0.8 \times 2.2\text{mm}$, and FOV = 212). We collected two identical runs, each with a single b_0 and 32 non-collinear gradient directions at $b = 700\text{s/mm}^2$. The data was pre-processed using the standard FDT (Jenkinson et al., 2012) pipeline in FSL consisting of susceptibility distortion correction, followed by corrections for eddy currents, motion and outliers. From here, tensor model fitting was performed to generate the transformation matrices and extract atlas based metrics. We used the BEDPOSTx tool in FSL (Behrens et al., 2007) to perform a bayesian estimation of the diffusion parameters at each voxel, followed by tractography using PROBTRACKx (Behrens et al., 2007).

Our experiments rely on the Automatic Anatomical Labelling (AAL) atlas (Tzourio-Mazoyer et al., 2002) parcellation for the rs-fMRI and DTI data. AAL consists of 116 cortical, subcortical and cerebellar regions. We employ a sliding window protocol as shown in Fig. 2 using the parameters learned in Section 2.2.1. Due to the different TR, we set the sliding window parameters to window length = 156 and stride = 17 for the HCP dataset, and window length = 45 and stride = 5 for the KKI dataset to extract dynamic correlation matrices from the 116 average time courses. We discuss the sensitivity to this choice in Section 4.2. Thus, for each individual, we have correlation matrices of size 116×116 based on the Pearson's Correlation Coefficient between the average regional time-series. Empirically, we observed a consistent noise component with nearly unchanging contribution from all brain regions and low predictive power for both datasets. Therefore, we subtracted out the first eigenvector contribution from each of the correlation matrices and used the residuals as the inputs $\{\mathbf{\Gamma}_n\}$ to the algorithm and the baselines.

Each DTI connectivity matrix \mathbf{A}_n is binary, where $[\mathbf{A}_n]_{ij} = 1$ corresponds to the presence of at least one tract between the regions i and j , 116 in total for AAL. For the KKI dataset, we impute the DTI connectivity for the 11 individual, who do not have DTI based on the training data in each cross validation fold.

Behavioral Data: For the HCP database, we examine the Cognitive Fluid Intelligence Score (CFIS) described in Duncan (2005) and Bilker et al. (2012), adjusted for age. This is scored based on a battery of

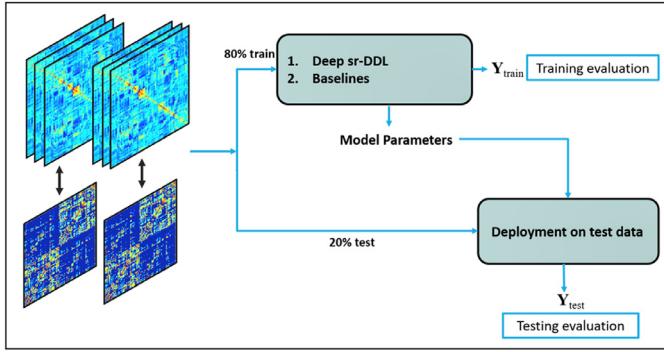


Fig. 8. A five-fold cross validation for evaluating performance.

tests measuring cognitive reasoning, considered a nonverbal estimate of fluid intelligence in subjects. The dynamic range for the score is 70–150, with higher scores indicating better cognitive abilities.

We analyzed three independent measures of clinical severity for the KKI dataset. These include:

- 1 Autism Diagnostic Observation Schedule, Version 2 (ADOS-2) total raw score
- 2 Social Responsiveness Scale (SRS) total raw score
- 3 Praxis total percent correct score

The ADOS consists of several sub-scores which quantify the social-communicative deficits in individuals along with the restrictive/repetitive behaviors (Lord et al., 2000). The test evaluates the child against a set of guidelines and is administered by a trained clinician. We compute the total score by adding the individual sub-scores. The dynamic range for ADOS is between 0 and 30, with higher score indicating greater impairment.

The SRS scale quantifies the level of social responsiveness of a subject (Bölte et al., 2008). Typically, these attributes are scored by parent/care-giver or teacher who completes a standardized questionnaire that assess various aspects of the child's behavior. Consequently, SRS reporting tends to be more variable across subjects, as compared to ADOS, since the responses are heavily biased by the parent/teacher attitudes. The SRS dynamic range is between 70 and 200 for ASD subjects, with higher values corresponding to higher severity in terms of social responsiveness.

Finally, Praxis is assessed using the Florida Apraxia Battery (modified for children) (Mostofsky et al., 2006). It assesses the ability to perform skilled motor gestures on command, by imitation, and with actual tool use. Several studies (Dowell et al., 2009; Dziuk et al., 2007; Mostofsky et al., 2006; Nebel et al., 2016) reveal that children with ASD show marked impairments in Praxis a.k.a., developmental dyspraxia, and that impaired Praxis correlates with impairments in core autism social-communicative and behavioral features. Performance is videotaped and later scored by two trained research-reliable raters, with total percent correctly performed gestures as the dependent variable of interest. Scores therefore range from 0–100, with higher scores indicating better Praxis performance. This measure was available for only 48 of the 57 subjects in the KKI dataset.

3.3. Evaluating predictive performance

We characterize the performance of each method using a five-fold cross validation strategy, as illustrated in Fig. 8.

We report three quantitative measures of performance. The first is the Median Absolute Error (MAE) between the outputs \hat{y}_n and the true scores y_n , computed as :

$$\text{MAE} = \text{median}(|\hat{y}_{:,m} - y_{:,m}|), \quad (14)$$

The MAE quantifies the absolute distance between the measured and predicted scores across individuals. We report MAE along with the corresponding standard deviation of the errors to quantify robustness. Lower MAE indicates better testing performance.

The second metric is the Normalized Mutual Information (NMI), which assesses the similarity in the distribution of the predicted and observed score distributions across subjects. NMI for the score m is computed as:

$$\text{NMI}(y_{:,m}, \hat{y}_{:,m}) = \frac{H(y_{:,m}) + H(\hat{y}_{:,m}) - H(y_{:,m}, \hat{y}_{:,m})}{\min \{H(y_{:,m}), H(\hat{y}_{:,m})\}} \quad (15)$$

Here, $H(y_{:,m})$ is the entropy of $y_{:,m}$ and $H(y_{:,m}, \hat{y}_{:,m})$ is the joint entropy between $y_{:,m}$ and $\hat{y}_{:,m}$. NMI ranges between 0–1 with a higher value indicating better agreement between predicted and measured score distributions, and thus characterizing improved performance.

Finally, we report the R^2 metric or the coefficient of determination evaluated on the predicted and true scores. Intuitively, the R^2 is a statistical measure that helps us assess the amount of variance in the true scores, i.e. y_m (for the m^{th}) score that is explained by the corresponding \hat{y}_m as predicted by the method. This is mathematically reported as

$$R^2(y_m, \hat{y}_m) = 1 - \frac{\sum_i (y_m(i) - \bar{y}_m)^2}{\sum_i (y_m(i) - \hat{y}_m(i))^2}$$

where, \bar{y}_m indicates the mean value of the true scores y_m . Larger values of R^2 indicate better agreement between the true and predicted scores.

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test	R^2 Test
CFIS	Median	N/A	13.51 ± 9.97	N/A	0	$1e^{-21}$
BC & LSTM-ANN		7.23 ± 6.24	16.50 ± 13.60	0.53	0.72	0.013
ICA & LSTM-ANN		4.87 ± 4.84	16.45 ± 14.7	0.58	0.77	0.013
BrainNet CNN		3.50 ± 2.1	16.89 ± 12.20	0.79	0.73	0.0017
Decoupled		3.72 ± 4.33	18.10 ± 14.04	0.78	0.70	0.011
Without DTI regularization		0.77 ± 0.66	20.02 ± 15.04	0.88	0.74	0.0089
Deep sr-DDL		0.44 ± 0.15	14.76 ± 12.77	<u>0.86</u>	0.77	0.071

3.4. Multi-Score prediction on real world data

Similarly, Fig. 9 illustrates the performance comparison of our deep sr-DDL framework against the baselines in Section 2.3 on the HCP dataset for predicting the CFIS. Fig. 10 presents the same comparison on the KKI dataset for multi-score prediction. In each figure, the scores predicted by the algorithm are plotted on the y-axis against the measured ground truth score on the x-axis. The bold $x = y$ line represents ideal performance. The red points represent the training data, while the Purple points indicate the held out testing data for all the cross validation folds.

We observe that the training performance of the baselines is good (i.e. the red points follow the $x = y$ line) in all cases for both datasets. However, in case of testing performance, our method outperforms the baselines in all cases. This performance gain is particularly pronounced in the case of multiscore prediction (KKI dataset). Empirically, we are able to tune the baseline hyperparameters to obtain good testing performance on the KKI dataset for a single score (ADOS for ICA + LSTM-ANN), but the prediction of the remaining scores (SRS and Praxis for the KKI dataset) suffers. Notice that the prediction on one or more of scores (KKI dataset) and CFIS (HCP dataset) hovers around the population median of the score in several cases. In fact, in some of the multi-score prediction cases, it performs worse than predicting the median. This is testament to the inherent difficulty of the prediction task at hand. Finally, we notice that omitting the structural regularization from the deep sr-DDL performs worse than our method.

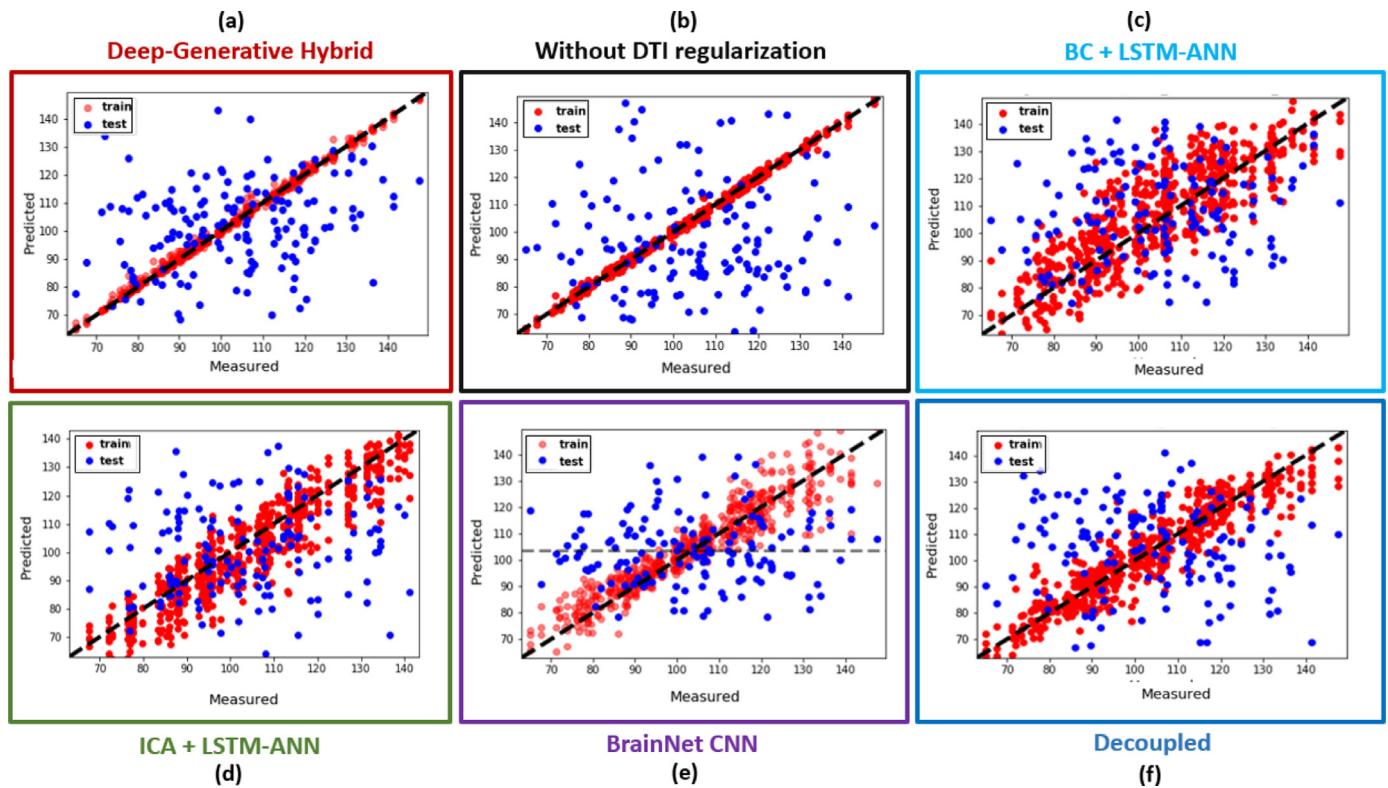


Fig. 9. HCP dataset: Prediction performance for the Cognitive Fluid Intelligence Score by the (a) Red Box: Deep sr-DDL. (b) Black Box: Deep sr-DDL model without DTI regularization (c) Light Purple Box: Betweenness Centrality on DTI + dynamic rs-fMRI multimodal graphs followed by LSTM-ANN predictor (d) Green Box: ICA timeseries followed by LSTM-ANN predictor (e) Purple Box: Branched BrainNet CNN (Kawahara et al., 2017) on DTI and rs-fMRI static graphs (f) Blue Box: Decoupled DDL factorization followed by LSTM-ANN predictor.

In contrast to the baselines, the testing predictions of our framework follow the $x = y$ more closely. The machine learning, statistical and graph theoretic techniques we selected for a comparison are well known in literature for being able to robustly provide compact characterizations for high dimensional datasets. However, we see that ICA is unable to estimate a reliable projection of the data that is particularly useful for behavioral prediction. Similarly, the betweenness centrality measure is unable to extract informative topologies for brain-behavior integration. We conjecture that the aggregate nature of this measure is useful for capturing group-level commonalities, but falls short of modeling subject-specific differences. Furthermore, even the BrainNet CNN, which directly exploits the graph structure of the connectomes falls short of generalizing to multi-score prediction. Additionally, it ignores the dynamic information in the rs-fMRI data. In case of the baseline where we omit the structural regularization, i.e. deep sr-DDL without DTI, we notice that the method learns a representation of the rs-fMRI data that generalizes beyond the training set, but still falls short of the performance when anatomical information is included. This clearly demonstrates the benefit of supplementing the functional data with structural priors. Finally, the failure of the decoupled dynamic matrix factorization and deep-network makes a strong case for jointly optimizing the neuroimaging and behavioral representations. The basis estimated independently of behavior are not indicative of clinical outcomes, due to which the regression performance suffers. We also quantify the performance indicated in these figures in Table 1 (HCP dataset) and Table 2 (KKI dataset) based on the MAE and NMI/R^2 . For reference, we have added an additional row as a ‘baseline’ in our tables where for each test subject, we simply predict the median of each score.

Our deep sr-DDL framework explicitly optimizes for a viable trade-off between multimodal and dynamic connectivity structures and behavioral data representations jointly. The dynamic matrix decomposition simultaneously models the group information through the basis, and the

subject-specific differences through the time-varying coefficients. The DTI Laplacians streamline this decomposition to focus on anatomically informed functional pathways. The LSTM-ANN directly models the temporal variation in the coefficients, with its weights encoding representations closely interlinked with behavior. The limited number of basis elements help provide compact representations explaining the connectivity information well. The regularization and constraints ensure that the problem is well posed, yet extracts clinically meaningful representations.

3.5. Clinical interpretation

Subnetwork Identification: In this section, we investigate the subnetworks learned in the basis **B** by the sr-DDL model when trained on both datasets. Recall that each column of the basis consists of a set of co-activated AAL subregions. In order to robustly identify these patterns, we first train the model on 10 randomly sampled subsets of each dataset. Then, we match the obtained subnetworks based on their absolute cosine similarity. Since we have 15 subnetworks, we then illustrate the mean co-activations across the brain regions for each of them individually in Fig. 11 (HCP) and Fig. 12 (KKI). Here, the colorbar in the figure indicates subnetwork contribution to the AAL regions. Regions storing negative values (cold colors) are anticorrelated with regions storing positive ones (hot colors). Alongside, we represent the corresponding standard deviations across different regions for each of the 15 subnetworks.

Examining the subnetworks in Fig. 11, we notice that Subnetworks 1 & 2, and 11 exhibits positive and competing contributions from regions of the Default Mode Network (DMN), which has been widely inferred in the resting state literature (Raichle, 2015) and is believed to play a critical role in consolidating memory (Sestieri et al., 2011), as also in self-referencing and in the theory of mind (Andrews-Hanna, 2012). At the same time, Subnetworks 2 and 11 have competing and positive contribu-

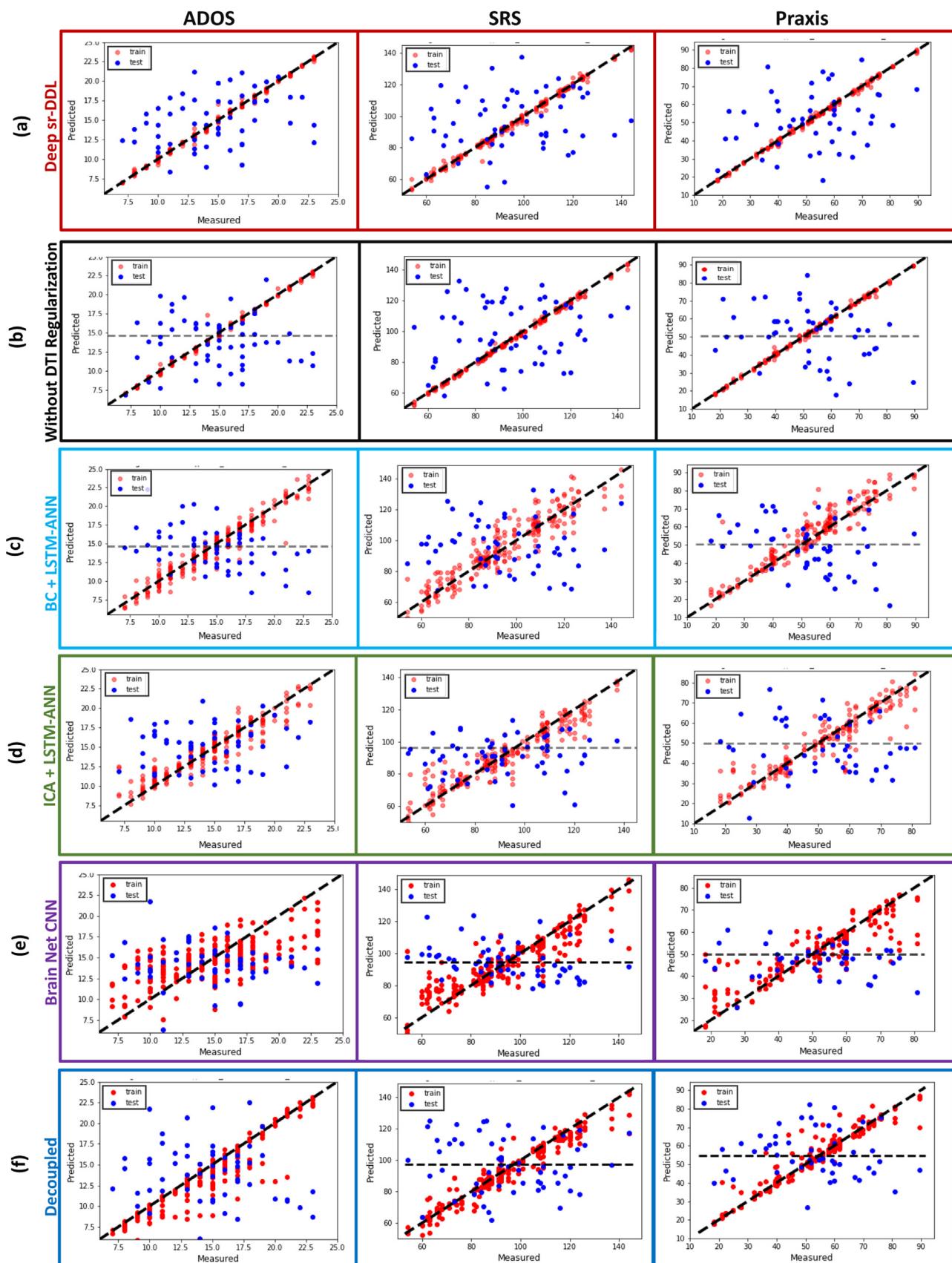


Fig. 10. KKI dataset: Multiscore prediction performance for the (L) ADOS, (M) SRS, and (R) Praxis by the (a) Red Box: Deep sr-DDL (b) Black Box: Model without DTI regularization (c) Light Purple Box: Betweenness Centrality on DTI + dynamic rs-fMRI multimodal graphs followed by LSTM-ANN predictor (d) Green Box: ICA timeseries followed by the LSTM-ANN predictor (e) Purple Box: Branched BrainNet CNN (Kawahara et al., 2017) on DTI Laplacian and rs-fMRI static graphs (f) Blue Box: Decoupled DDL factorization followed by LSTM-ANN predictor.

Table 1

KKI Dataset: Performance evaluation on the KKI dataset against our prior work according to **Median Absolute Error (MAE)**, **Normalized Mutual Information (NMI)**, and R^2 . We also report the standard deviation for the MAE. Lower MAE and higher NMI/ R^2 score indicate better performance. Best performance is highlighted in bold.

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test	R^2 Test
ADOS	Median	N/A	2.33 ± 2.01	N/A	0	$1e^{-31}$
	BC & LSTM-ANN	0.68 ± 0.57	4.36 ± 3.36	0.89	0.29	0.01
	ICA & LSTM-ANN	0.9 ± 0.54	2.47 ± 2.04	0.91	0.41	0.25
	BrainNet CNN	1.90 ± 0.086	3.50 ± 2.20	0.96	0.25	0.17
	Decoupled	1.34 ± 0.51	3.93 ± 2.10	0.68	0.29	0.06
	Without DTI regularization	0.25 ± 0.099	3.50 ± 3.09	0.99	0.17	0.02
	Deep sr-DDL	0.2 ± 0.09	2.99 ± 1.99	0.99	0.37	0.23
	Median	N/A	16.81 ± 12.8	N/A	0	$1e^{-30}$
	BC & LSTM-ANN	5.10 ± 4.61	18.05 ± 14.22	0.92	0.83	0.09
	ICA & LSTM-ANN	5.27 ± 3.32	13.64 ± 12.69	0.76	0.59	0.008
SRS	BrainNet CNN	5.25 ± 2.5	18.96 ± 15.65	0.83	0.75	0.018
	Decoupled	2.10 ± 2.98	21.45 ± 13.73	0.76	0.78	0.002
	Without DTI regularization	0.72 ± 0.61	22.20 ± 14.78	0.95	0.65	0.08
	Deep sr-DDL	1.21 ± 0.66	18.70 ± 13.51	0.98	0.85	0.12
	Median	N/A	10.53 ± 8.81	N/A	0	$1e^{-29}$
	BC & LSTM-ANN	6.61 ± 3.30	17.49 ± 9.08	0.86	0.70	0.01
Praxis	ICA & LSTM-ANN	4.56 ± 1.26	15.02 ± 11.80	0.82	0.60	0.0122
	BrainNet CNN	3.78 ± 0.59	15.15 ± 11.49	0.95	0.19	0.009
	Decoupled	1.57 ± 1.12	21.67 ± 12.02	0.75	0.25	0.003
	Without DTI regularization	0.61 ± 0.29	18.56 ± 14.32	0.96	0.65	0.08
	Deep sr-DDL	0.62 ± 0.36	14.99 ± 10.17	0.95	0.82	0.10

Table 2

Testing performance (5-fold CV) of the sr-DDL framework for single-target and multi-target prediction on the KKI dataset according to **Median Absolute Error (MAE)**, **Normalized Mutual Information (NMI)**, and R^2 . We also report the standard deviation for the MAE. Lower MAE and higher NMI/ R^2 scores indicate better performance.

Score	Method	MAE	NMI	R^2
ADOS	Single-target	2.91 ± 2.71	0.44	0.041
	Multi-target	2.99 ± 1.99	0.37	0.23
SRS	Single-target	14.78 ± 14.24	0.87	0.13
	Multi-target	18.70 ± 13.51	0.85	0.12
Praxis	Single-target	12.40 ± 11.60	0.85	0.06
	Multi-target	14.99 ± 10.17	0.82	0.10

tions from regions in the Frontoparietal Network (FPN) respectively. The FPN is known to be involved in executive function and goal-oriented, cognitively demanding tasks (Uddin et al., 2019). Subnetworks 1, 6, 7, 11 and 13 are comprised of regions from the Medial Frontal Network (MFN). The MFN and FPN are known to play a key role in decision making, attention and working memory (Euston et al., 2012; Menon, 2011), which are directly associated with cognitive intelligence. Subnetworks 1, 3, and 9 include contributions from the subcortical and cerebellar regions, while Subnetworks 10, 2, 14 and 11 include contributions from the Somatomotor Network (SMN). Taken together, these networks are believed to be important functional connectivity biomarkers of cognitive intelligence and consistently appear in previous literature on the HCP dataset (Chén et al., 2019; Hearne et al., 2016).

For the KKI dataset, in Fig. 12, Subnetwork 1 includes regions from the DMN, and the SMN. Similarly, Subnetwork 6 includes competing contributions from the SMN and DMN regions. Aberrant connectivity within the DMN and SMN regions have previously been reported in ASD (Lynch et al., 2013; Nebel et al., 2016). Subnetworks 7, 3, and 6 exhibit contributions from higher order visual processing areas in the occipital and temporal lobes along with competing sensorimotor regions. At the same time, Subnetwork 9 exhibits competing contributions from the visual network. These findings concur with behavioral reports of reduced visual-motor integration in autism (Nebel et al., 2016). Subnetworks 11 and 8 exhibit contributions from the central executive control

network (CEN) and insula. Subnetwork 10 also exhibits anticorrelated CEN contributions. These regions are believed to be essential for switching between goal-directed and self-referential behavior (Sridharan et al., 2008). Subnetwork 5 and Subnetwork 3 includes prefrontal and DMN regions, along with subcortical areas such as the thalamus, amygdala and hippocampus. The hippocampus is known to play a crucial role in the consolidation of long and short term memory, along with spatial memory to aid navigation. Altered memory functioning has been shown to manifest in children diagnosed with ASD (Williams et al., 2006). The thalamus is responsible for relaying sensory and motor signals to the cerebral cortex in the brain and has been implicated in autism-associated sensory dysfunction, a core feature of ASD (Cascio et al., 2008). Along with the amygdala, which is known to be associated with emotional responses, these areas may be crucial for social-emotional regulation in ASD. Pouw et al. (2013).

Finally, we notice that the standard deviations for a majority of the regions in each of the subnetworks are small compared to the mean coactivation. Additionally, we observed an average similarity of 0.79 ± 0.13 and 0.81 ± 0.12 for these subnetworks across the runs on subsets of the HCP and KKI datasets respectively. These results suggests that our deep-generative framework is able to capture stable underlying mechanisms which robustly explain the different sets of deficits in ASD as well as robustly extract signatures of cognitive flexibility in neurotypical individuals.

Study of Emerging Patterns:

In this experiment, we study the overlap in the subnetworks in the basis **B** across different scales of subnetworks, i.e. varying the number of networks K . Recall from Section 2.2.1, that the knee point of the eigen-spectrum of $\{\Gamma_n^t\}$ for both datasets is between 8 – 20. Namely, we re-run the sr-DDL model on both the datasets steadily increasing the number of networks from 8 – 20. In each case, we repeat the experiment using 10 random subsets of the data and look for subnetworks that appear most often. Figs. 11 and 12 illustrate the top ten networks that appear most frequently across different data subsets and choice of K for the HCP dataset and KKI dataset respectively. Alongside, we also report the mean and standard deviation of the absolute cosine similarity (S) for each individual subnetworks across the multiple runs. Networks which are most consistent exhibit higher similarity across runs with group 1 being the top five subnetworks ($S \geq 0.95$), group 2 being the next five subnetworks ($S > 0.85$). Finally, a visual inspection and comparison with our results

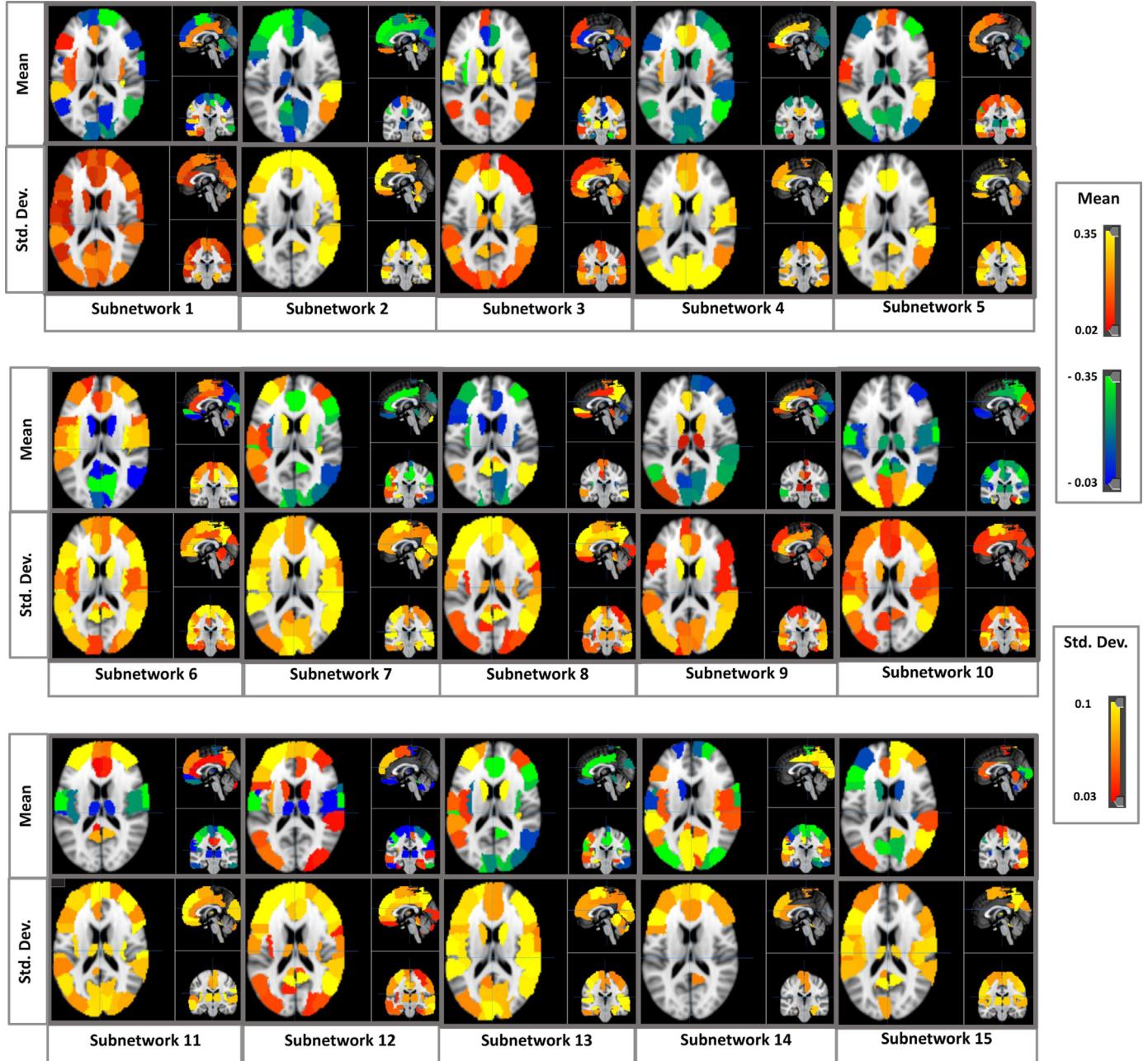


Fig. 11. Complete set of subnetworks identified by the deep sr-DDL model for the HCP database. **Mean:** Mean regional co-activation patterns in basis B. The red and orange regions are anti-correlated with the Purple and green regions. **Std. Dev.:** Standard deviations of regional co-activation patterns. A majority of regions exhibit small deviations from the mean. Both sets of plots have been computed across cross-validation folds

in Section 3.5 suggest a considerable overlap between the subnetworks in Figs. 11 and 13 for the HCP dataset and between Figs. 12 and 14 for the KKI dataset. These results suggest that our Deep sr-DDL robustly extracts representative neural signatures indicative of behavior in both healthy and autistic populations.

Decoding rs-fMRI networks dynamics:

Our deep sr-DDL allows us to map the evolution of functional networks in the brain by probing the LSTM-ANN representation. Recall that our model does not require the rs-fMRI scans to be of equal length. Fig. 15 (left) illustrates the learned attentions output by the A-ANN for the 150 subjects from the HCP dataset on the top and the 57 KKI subjects at the bottom during testing. For the KKI dataset, the patients with shorter scans have been grouped in the top of the figure. These time-points have been blackened at the beginning of the scan. The col-

orbar indicates the strength of the attention weights. Higher attention weights denote intervals of the scan considered especially relevant for prediction. Notice that the network highlights the start of the scan for several individuals, while it prefers focusing on the end of the scan for some others, especially pronounced in case of the KKI dataset. The patterns are comparatively more diffused for subjects in the HCP dataset, although several subjects manifest selectivity in terms of relevant attention weights. This is indicative of the underlying individual-level heterogeneity in both the cohorts.

Next, we illustrate the variation of the network strength for a representative subject from the HCP dataset and KKI dataset over the scan duration in Fig. 15 (right) at the top and bottom respectively. Each solid colored line corresponds to one of the 15 sub-networks in Fig. 12. Notice that, over the scan duration, each network cycles through phases

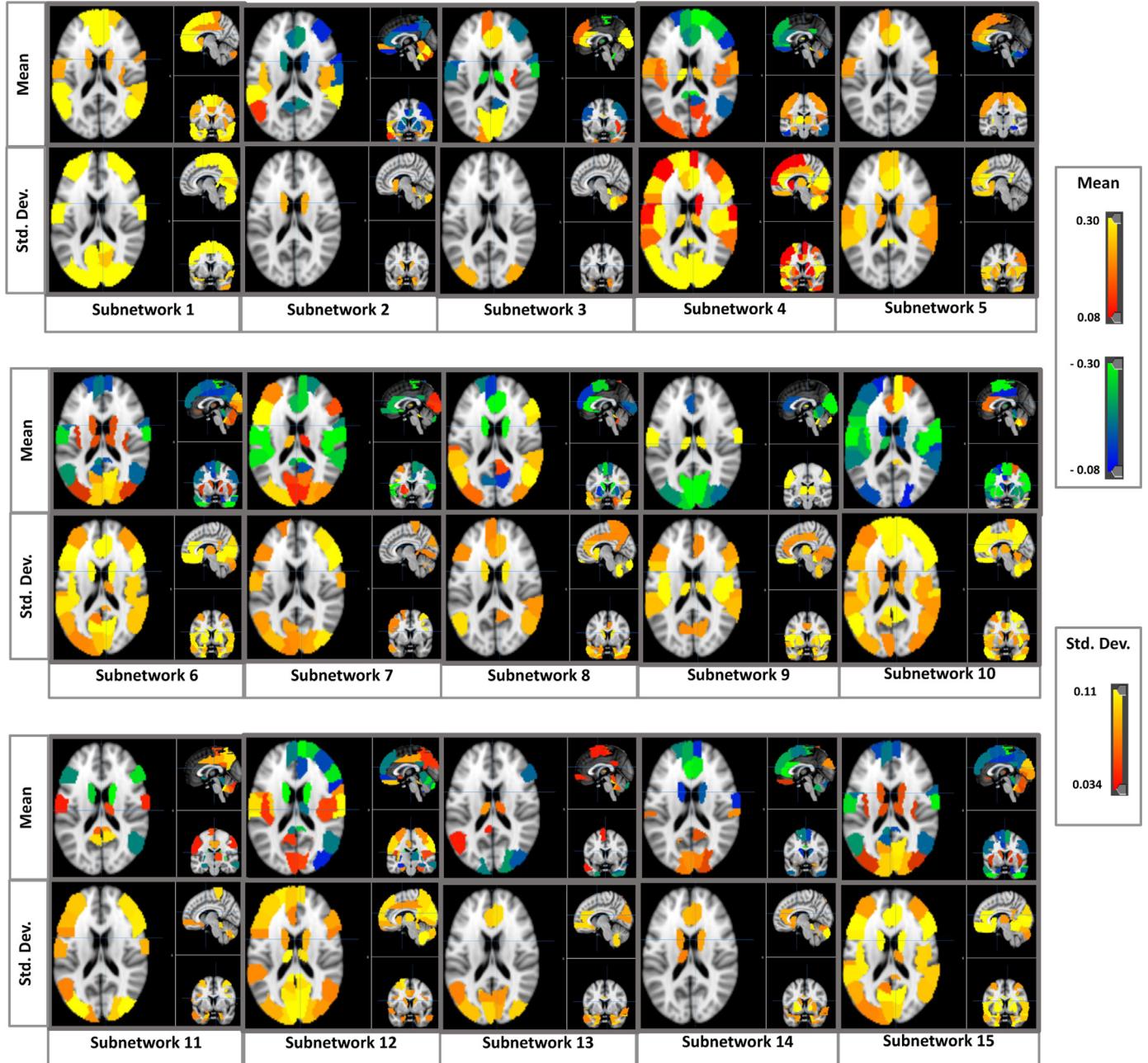


Fig. 12. Complete set of subnetworks identified by the deep sr-DDL model for the KKI database. **Mean:** Mean regional co-activation patterns in basis **B**. The red and orange regions are anti-correlated with the Purple and green regions. **Std. Dev.:** Standard deviations of regional co-activation patterns. A majority of regions exhibit small deviations from the mean. Both sets of plots have been computed across cross-validation folds

of activity and relative inactivity. Consequently, only a few networks at each time step contribute to the patient's dynamic connectivity profile. This parallels the transient brain-states hypothesis in dynamic rs-fMRI connectivity (Allen et al., 2014), with active states as corresponding sub-networks in the basis matrix **B**.

4. Discussion

Our deep-generative hybrid cleverly exploits the intrinsic structure of the rs-fMRI correlation matrices through the dynamic dictionary representation to simultaneously capture group-level and subject-specific information. At the same time, the LSTM-ANN network models the temporal evolution of the rs-fMRI data to predict behavior. The compactness of our representation serves as a dimensionality reduction step that is

related to the clinical score of interest, unlike the pipelined treatment commonly found in the literature. Our structural regularization helps us fold in anatomical information to guide the functional decomposition. Overall, our framework outperforms a variety of state-of-the-art graph theoretic, statistical and deep learning baselines on two separate real world datasets.

We conjecture that the baseline techniques fail to extract representative patterns from structural and functional data. These techniques are quite successful at modelling group level information, but fail to generalize to the entire spectrum of cognitive, symptomatic or connectivity level differences among subjects. Consequently, they overfit the training data.

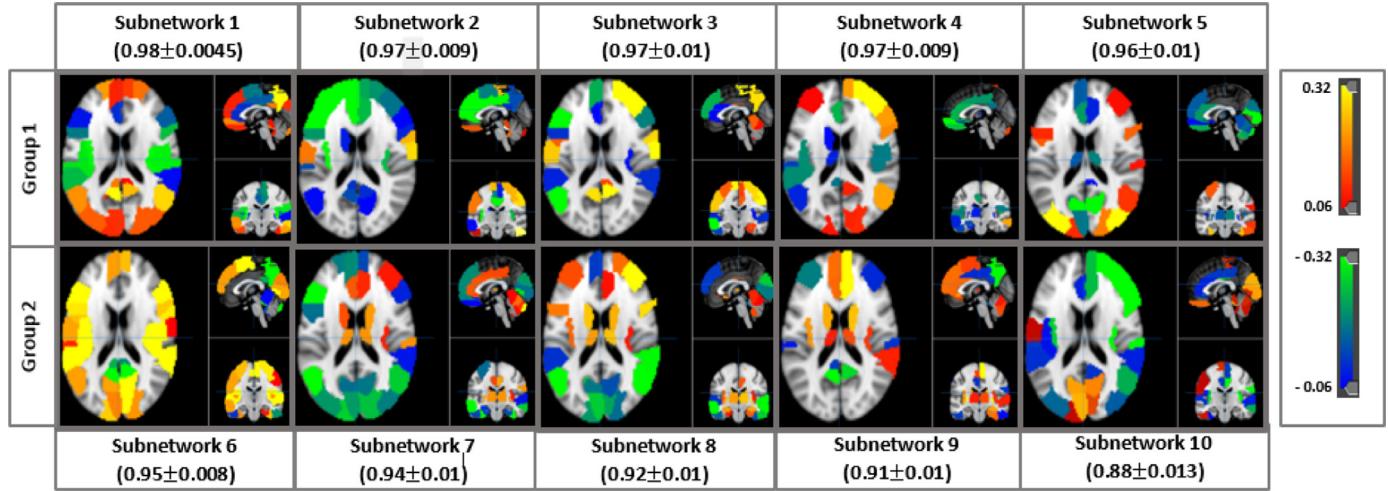


Fig. 13. HCP dataset: Set of top 10 consistent subnetworks across different model orders. Subnetworks in group 1 exhibit above 0.95 average similarity across data subsets and model orders. Subnetworks in group 2 exhibit between 0.85 – 0.95 average similarity across data subsets and model orders.

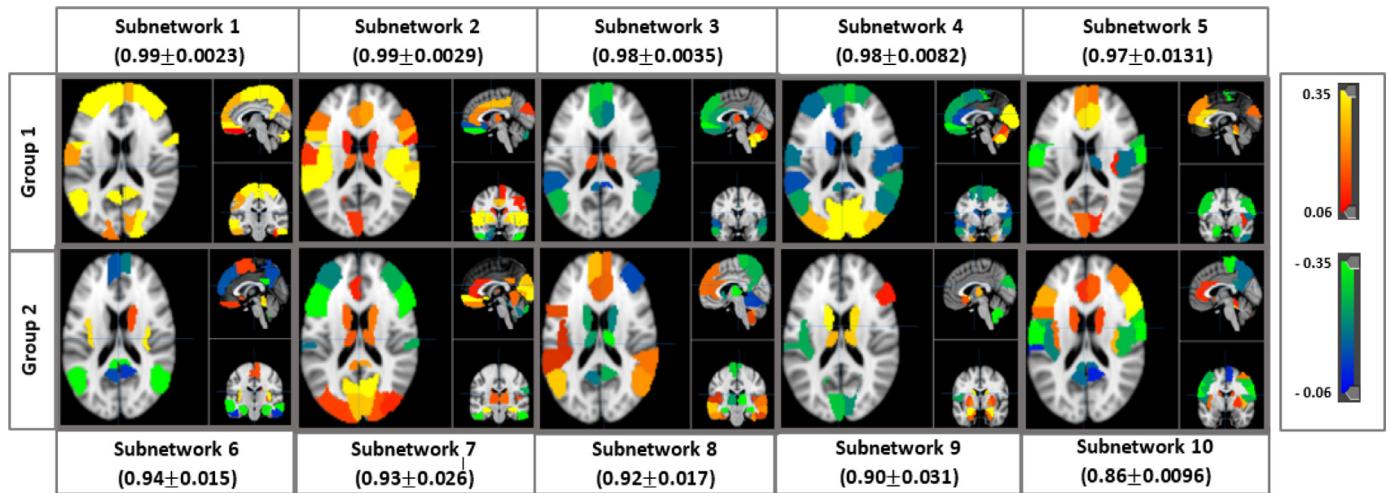


Fig. 14. KKI dataset: Set of top 10 consistent subnetworks across different model orders. Subnetworks in group 1 exhibit above 0.95 average similarity across data subsets and model orders. Subnetworks in group 2 exhibit between 0.85 and 0.95 average similarity across data subsets and model orders.

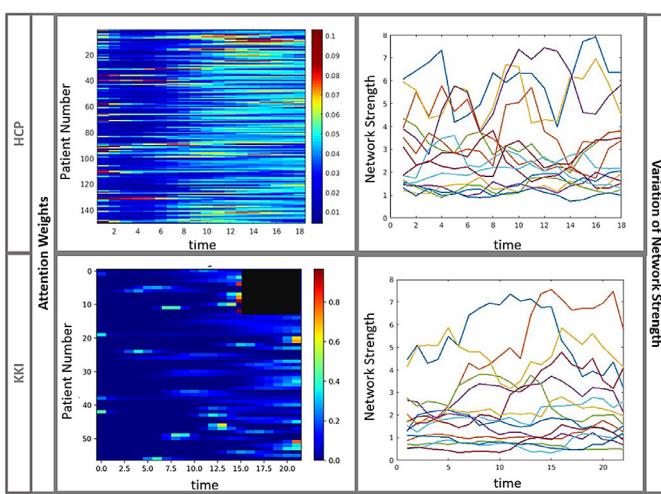


Fig. 15. (Left) Learned attention weights **(Right)** Variation of network strength over time on the **(Top)** HCP dataset **(Bottom)** KKI dataset.

4.1. Examining generalizability

Notice that the training examples (red points) in Figs. 9 and 10 follow the $x = y$ line perfectly, which may suggest overfitting. This phenomenon can be explained by the difference between our training procedure, where we optimize our joint objective in Eq. (8) assuming the scores are known, and our testing procedure. Recall that Section 2.2 describes the procedure for calculating the temporal sr-DDL loadings for an unseen patient i.e. \hat{c}_n^t from the basis \mathbf{B}^* obtained during training. Since the subject is not a part of the training set, the corresponding value of \hat{y} is unknown. Effectively, we must set the contribution from the data term, i.e., the deep network loss $\mathcal{L}(\cdot)$ in Eq. (8) to 0. Here, we examine the effect of employing the same strategy to calculate the coefficients for the training patients. In essence, we estimate the corresponding severity \hat{Y} now excluding the deep network loss. Accordingly, Fig. 16 highlights the differences in training fit with and without this term included in estimating $\{c_n^t\}$ for the HCP dataset. Notice that in the latter, the training accuracy for the CFIS score has the same distribution as the testing points in Fig. 9. In contrast, inclusion of the deep network loss in our coupled optimization overparameterizes the search space of solutions for $\{c_n^t\}$ to yield a near perfect fit.

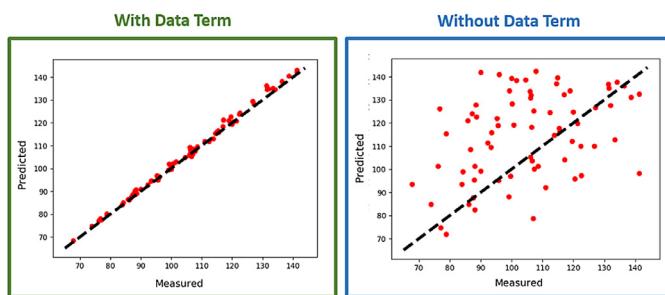


Fig. 16. Prediction Performance of the Deep sr-DDL for the CFIS score on training data when (L) The data term is included in computing $\{\mathbf{c}_n^t\}$ (R) The data term is excluded from the computation of $\{\mathbf{c}_n^t\}$.

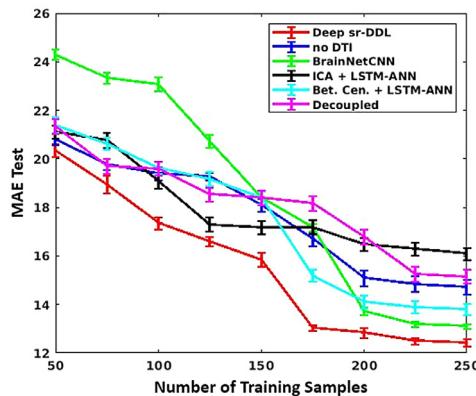


Fig. 17. Median Absolute Error on the Test Set varying the number of samples used for training. The vertical bars indicate standard errors for each setting.

To further probe the generalization capabilities of our Deep sr-DDL, we examine the effect of training the models on different sized datasets. For this experiment, we first set aside 50 individuals from the HCP database as a test set on which we evaluate the generalization performance. We then sweep the training set size from $N = 50 - 200$ in increments of 25 subjects. To avoid biasing the results, none of these subjects overlap with the HCP-2 validation set used for parameter tuning in Section 2.2.1. For each training set size, we randomly sample the subjects 10 times and compute the generalization performance on the held-out set.

Fig. 17 displays the MAE of the CFIS score prediction on the test set as a function of the training set size. As expected, we observe that with increasing training data, the performance on the test set improves at first but eventually saturates for all methods. This is evinced by a lowering of the MAE in the initial parts of the curve followed by a subsequent plateau at roughly 150–200 samples. Based on these results, we conjecture that further addition of training data does not substantially improve the generalization capabilities of our model or the baselines. We also note that the deep sr-DDL outperforms the baselines across the entire regime. In conjunction with our results from Section 3.2, we conclude that the deep sr-DDL model performs reasonably well for small to moderately sized datasets. This is especially important against the backdrop of potential clinical applications, many of which have datasets of modest sizes.

4.2. Assessing model robustness

Our deep sr-DDL framework has only two free hyperparameters. The first is the number of subnetworks in \mathbf{B} . As described in Section 2.2.1, we use the eigen-spectrum of $\{\mathbf{I}_n^t\}$ to fix this at 15 for both datasets. The second is the penalty parameter λ , which controls the trade-off between representation and prediction. Recall that our data pre-processing

includes a sliding window protocol in Fig. 2, which is defined by two parameters, i.e. the sliding window length and the stride. From a mathematical perspective, our deep sr-DDL formulation as such is agnostic to these parameters, as they are simply folded into the input data dimension. However, empirically, they balance the context size and information overlap within the rs-fMRI correlation matrices $\{\mathbf{I}_n^t\}$ and affects the prediction performance.

In this section, we evaluate the performance of our framework under three scenarios. Specifically, we sweep λ , the window length and the stride parameter independently, keeping the other two values fixed. We use five fold cross validation with the MAE metric to quantify the multi-score prediction performance, which as shown in Section 3.2, is more challenging than single score prediction. Fig. 18 plots the performance for the three scores on the KKI dataset with MAE value for each score on the y axis and the parameter value on the x axis.

We observed that our method gives stable performance for fairly large ranges of each parameter settings. As expected, low values of λ (0.01–1) result in higher MAE values, likely due to underfitting. Similarly, higher values >6 result in overfitting to the training dataset, degrading the generalization performance. Additionally, lower values of window lengths result in higher variance among the correlation values due to noise, and hence less reliable estimates of dynamic connectivity (Lindquist, 2016). On the other hand, very large context windows tend to miss nuances in the dynamic evolution of the scan. Empirically, we observe that a mid-range of window length 100–125s yields a good tradeoff between representation and prediction. The training of LSTM networks with very long sequence lengths is known to be particularly challenging owing to vanishing/exploding gradient issues during back-propagation. However, having too short a sequence confounds a reliable estimation of the LSTM weights from limited data. The stride parameter helps mitigate these issue by compactly summarizing the information in the sequence while simultaneously controlling the overlap across subsequent samples. Our experiments found a stride length between 10–20s to be suitable for our application.

In summary, the guidelines we identified for each of the parameters are- $\lambda \in (2 - 5)$, window length $\in (100 - 125)$ s, and stride $\in (10 - 20)$ s. Additionally, our experiments on the HCP dataset using the same settings indicate that the results of our method are reproducible across different populations. It is also interesting to note that previous experiments on the HCP dataset in literature have found similar window lengths to be stable in classification (Gadgil et al., 2020) and various test-retest settings (Savva et al., 2019).

4.3. Clinical relevance

Our experiments on the KKI dataset evaluate the ability of our Deep sr-DDL framework to simultaneously explain multiple clinical impairments of ASD. This multi-target prediction is a challenging task, and in fact, the baseline methods fail to generalize all three scores. At the same time, one could evaluate the performance of predicting each score independently via three single-target regression tasks. Accordingly, Table 2 compares the performance of our Deep sr-DDL framework in the single-target and multi-target settings. Empirically, we observe that the single-target prediction is slightly better than the multi-target prediction. Indeed, a possible counter perspective would be to optimize for prediction accuracy of individual measures explained by potentially different brain bases, for example, as in the work of D'Souza et al. (2019a). This comparison poses a more philosophical question about the benefits of a multi-target setup given a possible decline in predictive performance and the difficulty of the task itself.

To weigh in on this trade off, we note the growing consensus in clinical psychiatry that complex disorders, such as autism and schizophrenia, are inherently multidimensional (Havdahl et al., 2016). Furthermore, there is considerable patient heterogeneity within a single diagnostic umbrella that reflect subtle differences in the underlying etiology (Hong et al., 2018). In fact, the National Institute of Mental Health

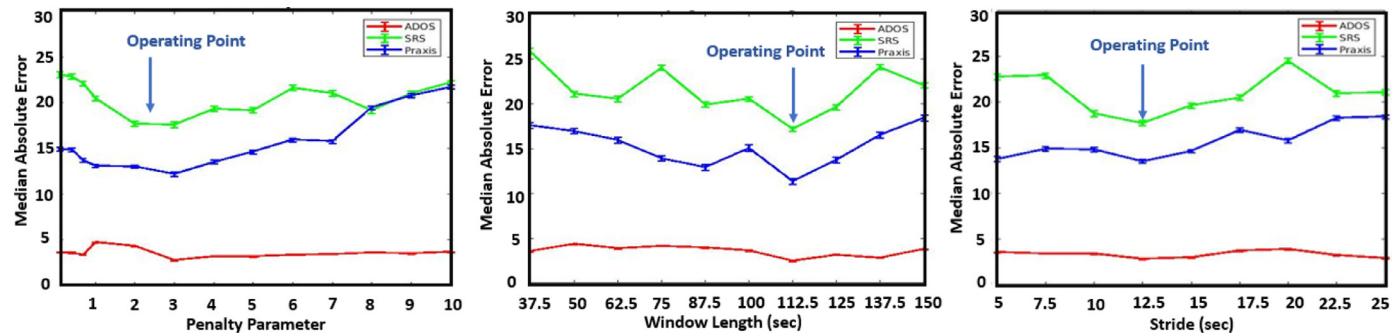


Fig. 18. Performance of the Deep sr-DDL upon varying (L): the penalty parameter λ (B): window length (R): stride. Our operating point is indicated by the Purple arrow.

(NIHM) in the United States has released the RDoc research framework (Insel, 2014), which advocates for a multidimensional characterization to understand the full spectrum of mental health and illness. In this context, our Deep sr-DDL approach provides a flexible tool to map multiple measures via a consistent and stable brain basis (as shown by the results in Section 3.5). Thus, we view it as an important foundation to parse complex spectrum disorders that may even spur new analytical directions in brain connectomics.

Finally, our Deep sr-DDL framework is carefully designed to extract subject-level dynamic information. Namely, the attention mechanism automatically highlights portions of the rs-fMRI scan that are important for clinical prediction (Fig. 15). In fact, a comparison of the attention weights in Fig. 15 suggests considerable inter-patient variability of the intervals used for multi-target prediction in the KKI dataset, as opposed to the relatively consistent attention weights in the HCP dataset. This pattern may be linked to the heterogeneity of ASD described above. In conjunction, we observe the subnetwork contributions phasing in and out prominence over the course of the scan, which is consistent with the transient brain state hypothesis (Allen et al., 2014).

In summary, the blend of classical generative modeling and deep learning prediction in our Deep sr-DDL framework allows for a finer-grained characterization of connectivity and behavior. Overall, we believe that the robustness, stability, clinical interpretability, and flexibility of our Deep sr-DDL render it a novel and valuable tool for the research community.

4.4. Applications, limitations and future scope

As seen in our experiments in Section 3.4, our method is able to extract key predictive resting state biomarkers from healthy and autistic populations. Additionally, our deep sr-DDL makes minimal assumptions. Provided we have access to a set of consistently defined structural and functional connectivity measures and clinical scores, this analysis can be easily adapted to other neurological disorders and even predictive network models outside the medical realm. Overall, these findings broaden the scope of our method for future applications.

Although we outperform several baselines on two separate datasets, our prediction performance in Section 3.4 is far from perfect. This underscores that multi-score prediction is a challenging clinical problem. One of the key reasons can be attributed to inherent noise in the clinical measures themselves. For example, SRS is based on a parent-teacher questionnaire, which tends to be more subjective than a clinical exam. This renders the behavioral prediction task especially challenging, which partially accounts for the poor performance of several baselines we compared against. Keeping this in mind, a natural clinical direction of exploration is to adopt our method to predicting measures more directly related to functional connectivity, as opposed to those relying on clinical reports. Another avenue of exploration includes examining more coarse indicators of behavior, such as ordered levels of impairment instead of

continuous measures (an ordinal regression problem), or the prevalence of ASD sub-types.

Another limitation to our method lies in the fact that our estimate of dynamic functional connectivity relies on the availability of a reliable sliding-window protocol. As illustrated in Section 4.2, an inappropriate window-length and stride choice has a direct bearing on the predictive performance. Moreover, this tradeoff is difficult to quantify and correct for analytically. Keeping this in mind, we are motivated to explore alternatives to the sliding window for better estimating dynamic functional connectivity, which can at the same time be robustly integrated into multimodal data-analysis frameworks such as ours.

From the methodological standpoint, we recognize that our model is simplistic in its assumptions, particularly in the sr-DDL formulation. The DTI priors guide a data-driven classical rs-fMRI matrix decomposition in a regularization framework. This modeling choice was deliberately employed to preserve interpretability in the basis and simplify the inference procedure. A key limitation of this approach is that it does not directly consider multi-stage pathways, which may be an important mediator of functional relationships between communicating sub-regions. To this end, graph neural networks have shown great promise in brain connectivity research due to their ability to capture subtle and multi-stage interactions between communicating brain regions while exploiting the underlying hierarchy of brain organization. Consequently, these methods are emerging as important tools to probe complex pathologies in brain functioning and diagnose neurodevelopmental disorders (Anirudh and Thiagarajan, 2019; Parisot et al., 2018). In the future, we are exploring end-to-end graph convolutional networks that model the evolution of rs-fMRI signals on the anatomical DTI graphs.

5. Conclusion

We have introduced a novel deep-generative framework to integrate complementary information from the functional and structural neuroimaging domains, which simultaneously maps to behavior. Our unique structural regularization elegantly injects anatomical information into the rs-fMRI functional decomposition, thus providing us with an interpretable brain basis. Our deep network (LSTM-ANN) not only models the temporal variation among individuals, but also helps isolate key dynamic resting-state signatures, indicative of clinical/cognitive impairments. Our coupled optimization procedure ensures that we learn effectively from limited training data while generalizing well to unseen subjects. Finally, our framework makes very few assumptions and can potentially be applied to study other neuropsychiatric disorders (eg. ADHD, Schizophrenia) as an effective diagnostic tool.

Credit authorship contribution statement

N.S. D'Souza: Conceptualization, Methodology, Software, Formal analysis, Investigation, Validation, Writing – original draft, Writing – review & editing. **M.B. Nebel:** Conceptualization, Writing – review & edit-

ing, Funding acquisition, Project administration, Data curation, Supervision, Resources. **D. Crocetti:** Conceptualization, Project administration, Data curation, Resources. **J. Robinson:** Data curation, Resources. **N. Wymbs:** Data curation, Resources. **S.H. Mostofsky:** Conceptualization, Writing – review & editing, Funding acquisition, Project administration, Supervision, Resources. **A. Venkataraman:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition, Project administration.

Acknowledgements

This work has generously been supported by the [National Science Foundation](#) CRCNS award [1822575](#) and CAREER award [1845430](#), the [National Institute of Mental Health](#) (R01 MH085328-09, R01 MH078160-07, K01 MH109766 and R01 MH106564), the [National Institute of Neurological Disorders and Stroke](#) (R01NS048527-08), and the Autism Speaks foundation.

Appendix A

Here, we provide the detailed derivations for the Weighted Frobenius Norm expression in [Eq. \(4\)](#). We begin with the formulation in [Eq. \(3\)](#) below:

$$\|\Gamma_n^t - \mathbf{B}\text{diag}(\mathbf{c}_n^t)\mathbf{B}^T\|_{\mathbf{L}_n} = \|\mathbf{E}_n^t\|_{\mathbf{L}_n} \quad (\text{A.1})$$

Here, \mathbf{E}_n^t represents the reconstruction error in the correlation matrix Γ_n^t for patient n at time t . For the DTI graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for patient n , $\mathbf{L}_n = \mathbf{V}_n^{-\frac{1}{2}}(\mathbf{V}_n - \mathbf{A}_n)\mathbf{V}_n^{-\frac{1}{2}}$ is the DTI Graph Laplacian, where $\mathbf{V}_n = \text{diag}(\mathbf{A}_n\mathbf{1})$ is the degree matrix and $\mathbf{1}$ is the vector of all ones. For notational convenience, we will drop the subscripts n and t from the following computation.

$$\begin{aligned} \|\mathbf{E}\|_{\mathbf{L}} &= \text{Tr}[\mathbf{E}^T \mathbf{L} \mathbf{E}] = \text{Tr}[\mathbf{E}^T \mathbf{V}^{-\frac{1}{2}}(\mathbf{V} - \mathbf{A})\mathbf{V}^{-\frac{1}{2}} \mathbf{E}] \\ &= \text{Tr}[\tilde{\mathbf{E}}^T (\mathbf{V} - \mathbf{A}) \tilde{\mathbf{E}} \text{ where } \tilde{\mathbf{E}} = \mathbf{V}^{-\frac{1}{2}} \mathbf{E}] \\ &= \sum_i \sum_j \sum_k \tilde{\mathbf{E}}(i, j)[\mathbf{V}(i, k) - \mathbf{A}(i, k)]\tilde{\mathbf{E}}(k, j) \\ &= \sum_{i, j, k} \mathbf{V}(i, k)\tilde{\mathbf{E}}(i, j)\tilde{\mathbf{E}}(k, j) - \sum_{i, j, k} \mathbf{A}(i, k)\tilde{\mathbf{E}}(i, j)\tilde{\mathbf{E}}(k, j) \\ &= \sum_{i, j} \mathbf{V}(i, i)\tilde{\mathbf{E}}(i, j)\tilde{\mathbf{E}}(i, j) - \sum_{i, j, k} \mathbf{A}(i, k)\tilde{\mathbf{E}}(i, j)\tilde{\mathbf{E}}(k, j) \\ &= \sum_j \sum_{(i, k) \in \mathcal{E}} 2[\tilde{\mathbf{E}}(i, k)]^2 - \sum_j \sum_{(i, k) \in \mathcal{E}} 2[\tilde{\mathbf{E}}(i, j)\tilde{\mathbf{E}}(k, j)] \\ &= \sum_j \left[\sum_{(i, k) \in \mathcal{E}} [\tilde{\mathbf{E}}(i, k)]^2 + \sum_{(i, k) \in \mathcal{E}} [\tilde{\mathbf{E}}(k, j)]^2 \right] \\ &\quad - \sum_j \sum_{(i, k) \in \mathcal{E}} 2[\tilde{\mathbf{E}}(i, j)\tilde{\mathbf{E}}(k, j)] \\ &= \sum_j \sum_{(i, k) \in \mathcal{E}} [\tilde{\mathbf{E}}(i, j) - \tilde{\mathbf{E}}(k, j)]^2 \\ &= \sum_{(i, k) \in \mathcal{E}} \|\tilde{\mathbf{E}}(i, :) - \tilde{\mathbf{E}}(k, :)\|_2^2 \\ &= \sum_{(i, k) \in \mathcal{E}} \|[\mathbf{V}(i, i)]^{-\frac{1}{2}} \mathbf{E}(i, :) - [\mathbf{V}(k, k)]^{-\frac{1}{2}} \mathbf{E}(k, :)\|_2^2 \end{aligned}$$

Writing out the appropriate subscripts and superscripts we dropped earlier, we obtain the expression in [Eq. \(4\)](#):

$$\begin{aligned} \|\Gamma_n^t - \mathbf{B}\text{diag}(\mathbf{c}_n^t)\mathbf{B}^T\|_{\mathbf{L}_n} &= \sum_{(i, k) \in \mathcal{E}} \|\tilde{\mathbf{E}}_n^t(i, :) - \tilde{\mathbf{E}}_n^t(k, :)\|_2^2 \\ &= \sum_{(i, k) \in \mathcal{E}} \|[\mathbf{V}_n(i, i)]^{-\frac{1}{2}} \mathbf{E}_n^t(i, :) - [\mathbf{V}_n(k, k)]^{-\frac{1}{2}} \mathbf{E}_n^t(k, :)\|_2^2 \end{aligned}$$

$$- [\mathbf{V}_n(k, k)]^{-\frac{1}{2}} \mathbf{E}(k, :)\|_2^2$$

Appendix B

In this section, we detail the calculations from [Section 2.2](#). Thus, our alternating minimization steps are explained as:

Step 1: Closed form solution for \mathbf{B} : Notice that [Eq. \(9\)](#) reduces to the following quadratic form in \mathbf{B} :

$$\mathbf{B}^* = \arg \min_{\mathbf{B}: \mathbf{B}^T \mathbf{B} = \mathcal{I}_K} \|\mathbf{M} - \mathbf{B}\|_F^2 \quad (\text{B.1})$$

where \mathbf{M} is computed as:

$$\begin{aligned} \mathbf{M} &= \sum_n \frac{1}{T_n} \sum_t (\Gamma_n^t \mathbf{L}_n + \mathbf{L}_n \Gamma_n^t) \mathbf{D}_n^t + \\ &\quad \sum_n \frac{1}{T_n} \left[\sum_t \frac{\gamma}{2} \mathbf{D}_n^t \text{diag}(\mathbf{c}_n^t) + \gamma \Lambda_n^t \text{diag}(\mathbf{c}_n^t) \right] \end{aligned} \quad (\text{B.2})$$

We know that \mathbf{B} has a closed-form Procrustes solution ([Everson, 1998](#)) computed as follows. Given the singular value decomposition $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, we have:

$$\mathbf{B}^* = \mathbf{U}\mathbf{V}^T$$

In essence, \mathbf{B} spans the anatomically weighted space of subject-specific dynamic correlation matrices.

Step 2: Updating the sr-DDL loadings $\{\mathbf{c}_n^t\}$: The objective \mathcal{J}_c in [Eq. \(9\)](#) decouples across subjects. We can also incorporate the non-negativity constraint $\mathbf{c}_{nk}^t \geq 0$ by passing an intermediate vector $\hat{\mathbf{c}}_n^t$ through a ReLU. Thus:

$$\mathbf{c}_n^t = \text{ReLU}(\hat{\mathbf{c}}_n^t) \quad (\text{B.3})$$

The ReLU pre-filtering allows us to optimize an unconstrained version of [Eq. \(9\)](#), as follows:

$$\begin{aligned} \mathcal{J}_c &= \lambda \sum_n \mathcal{L}(\Theta, \{\mathbf{c}_n^t\}; \mathbf{y}_n) \\ &\quad + \sum_{n, t} \frac{\gamma}{T_n} \left[\text{Tr}[(\Lambda_n^t)^T (\mathbf{D}_n^t - \mathbf{B}\text{diag}(\mathbf{c}_n^t))] \right] \\ &\quad + \sum_{n, t} \frac{\gamma}{T_n} \left[\frac{1}{2} \|\mathbf{D}_n^t - \mathbf{B}\text{diag}(\mathbf{c}_n^t)\|_F^2 \right] \end{aligned} \quad (\text{B.4})$$

This optimization can be performed via the stochastic ADAM algorithm ([Kingma and Ba, 2015](#)) by backpropagating the gradients from the loss in [Eq. \(B.4\)](#) upto the input $\{\hat{\mathbf{c}}^t\}$. Experimentally, we set the initial learning rate to be 0.02, scaled by 0.9 per 10 iterations. Essentially, this optimization couples the parametric gradient from the Augmented Lagrangian formulation with the backpropagated gradient from the deep network (parametrized by fixed Θ). After convergence, the thresholded loadings $\mathbf{c}_n^t = \text{ReLU}(\hat{\mathbf{c}}_n^t)$ are used in the subsequent steps of the minimization.

Step 3: Updating the Deep Network weights- Θ : We use backpropagation on the loss $\mathcal{L}(\cdot)$ to solve for the unknowns Θ . Notice that we can handle missing clinical data by dropping the contributions of the unknown value of \mathbf{y}_{nm} to the network loss during backpropagation. Again, we use the ADAM optimizer ([Kingma and Ba, 2015](#)) with random initialization at the first main iteration of alternating minimization. We employ a learning rate of $0.2e^{-4}$, scaled by 0.95 every 5 epochs, and batch-size 1. Additionally, we train the network only for 60 epochs to avoid overfitting.

Step 4: Updating the Constraint Variables $\{\mathbf{D}_n^t, \Lambda_n^t\}$: Each of the primal variables $\{\mathbf{D}_n^t\}$ has a closed form solution given by:

$$[\mathbf{D}_n^t]^k = \mathbf{K}\mathbf{F} \quad (\text{B.5})$$

where, $\mathbf{K} = (\text{diag}(\mathbf{c}_n^t)\mathbf{B}^T + \Gamma_n^t \mathbf{L}_n \mathbf{B} + \mathbf{L}_n \Gamma_n^t \mathbf{B} - \gamma \Lambda_n^t)$ and $\mathbf{F} = (\gamma \mathcal{I}_K + 2\mathbf{L}_n)^{-1}$. We update the dual variables $\{\Lambda_n^t\}$ via gradient ascent:

$$[\Lambda_n^t]^{k+1} = [\Lambda_n^t]^k + \eta_k ([\mathbf{D}_n^t]^k - \mathbf{B}\text{diag}(\mathbf{c}_n^t)) \quad (\text{B.6})$$

We cycle through the primal-dual updates for $\{\mathbf{D}_n^t\}$ and $\{\Lambda_n^t\}$ in Eqs. [\(B.5\)](#) and [\(B.6\)](#) to ensure that the constraints $\mathbf{D}_n^t = \mathbf{B}\text{diag}(\mathbf{c}_n^t)$ are satisfied with

increasing certainty at each iteration. The learning rate parameter η_k for the gradient ascent step is selected to a guarantee sufficient decrease in the objective for every iteration of alternating minimization. In practice, we initialize η_0 to 10^{-3} , and scale it by 0.75 at each iteration k .

Step 5: Prediction on Unseen Data: In our cross-validated setting, we must compute the sr-DDL loadings $\{\bar{\mathbf{c}}^t\}_{t=1}^{\bar{T}}$ for a new subject based on the \mathbf{B}^* obtained from the training procedure and the new rs-fMRI correlation matrices $\{\bar{\Gamma}^t\}$ and DTI Laplacians $\bar{\mathbf{L}}$. As we do not know the score \bar{y} for this individual, we need remove the contribution $\mathcal{L}(\cdot)$ from Eq. (9) and assume that the constraints $\bar{\mathbf{D}}^t = \mathbf{B}^* \text{diag}(\bar{\mathbf{c}}^t)$ are satisfied with equality. This effectively eliminates the Lagrangian terms. Essentially, the optimization for $\{\bar{\mathbf{c}}^t\}$ now reduces to \bar{T}_n decoupled quadratic programming (QP) objectives \mathcal{Q}_t :

$$\begin{aligned} \bar{\mathbf{c}}^{*t} &= \arg \min_{\bar{\mathbf{c}}^t} \frac{1}{2} (\bar{\mathbf{c}}^t)^T \bar{\mathbf{H}} \bar{\mathbf{c}}^t + \bar{\mathbf{f}}^T \bar{\mathbf{c}}^t \quad \text{s.t.} \quad \bar{\mathbf{A}} \bar{\mathbf{c}}^t \leq \bar{\mathbf{b}} \\ \bar{\mathbf{H}} &= 2(\mathbf{B}^{*T} \bar{\mathbf{L}} \mathbf{B}^*); \\ \bar{\mathbf{f}} &= -[\mathcal{I}_K \circ (\mathbf{B}^{*T} (\bar{\Gamma}^t \bar{\mathbf{L}} + \bar{\mathbf{L}} \bar{\Gamma}^t) \mathbf{B}^*)] \mathbf{1}; \\ \bar{\mathbf{A}} &= -\mathcal{I}_K \bar{\mathbf{b}} = \mathbf{0} \end{aligned}$$

Where \circ is the elementwise Hadamard product. Notice that decoupling the objective across time allows us to parallelize this computation. Additionally, since $\bar{\mathbf{H}}$ is positive semi-definite, the formulation above is convex, leading to an efficient QP solution. Finally, we estimate \bar{y} via a forward pass through the LSTM-ANN.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.neuroimage.2021.118388](https://doi.org/10.1016/j.neuroimage.2021.118388)

References

Aghdam, M.A., Sharifi, A., Pedram, M.M., 2018. Combination of RS-fMRI and SMRI data to discriminate autism spectrum disorders in young children using deep belief network. *J. Digit. Imaging* 31 (6), 895–903.

Aielli, G.P., 2013. Dynamic conditional correlation: on properties and estimation. *J. Bus. Econ. Stat.* 31 (3), 282–299.

Allen, E.A., Damaraju, E., Plis, S.M., Erhardt, E.B., Eichele, T., Calhoun, V.D., 2014. Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex* 24 (3), 663–676.

Andrews-Hanna, J.R., 2012. The brain's default network and its adaptive role in internalmentation. *Neuroscientist* 18 (3), 251–270.

Andrews-Hanna, J.R., Snyder, A.Z., Vincent, J.L., Lustig, C., Head, D., Raichle, M.E., Buckner, R.L., 2007. Disruption of large-scale brain systems in advanced aging. *Neuron* 55 (5), 924–935.

Anirudh, R., Thiagarajan, J.J., 2019. Bootstrapping graph convolutional neural networks for autism spectrum disorder classification. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3197–3201.

Assaf, Y., Pasternak, O., 2008. Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. *J. Mol. Neurosci.* 34 (1), 51–61.

Atasoy, S., Donnelly, I., Pearson, J., 2016. Human brain networks function in connectome-specific harmonic waves. *Nat. Commun.* 7, 10340.

Bardella, G., Bifone, A., Gabrielli, A., Gozzi, A., Squartini, T., 2016. Hierarchical organization of functional connectivity in the mouse brain: a complex network approach. *Scientific reports* 6, 32060.

Bassett, D.S., Bullmore, E., 2006. Small-world brain networks. *Neuroscientist* 12 (6), 512–523.

Behrens, T.E., Berg, H.J., Jbabdi, S., Rushworth, M.F., Woolrich, M.W., 2007. Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage* 34 (1), 144–155.

Bilker, W.B., Hansen, J.A., Brensinger, C.M., Richard, J., Gur, R.E., Gur, R.C., 2012. Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment* 19 (3), 354–369.

Bölte, S., Poustka, F., Constantino, J.N., 2008. Assessing autistic traits: cross-cultural validation of the social responsiveness scale (SRS). *Autism Res.* 1 (6), 354–363.

Bowman, F.D., Zhang, L., Derado, G., Chen, S., 2012. Determining functional connectivity using fMRI data with diffusion-based anatomical weighting. *NeuroImage* 62 (3), 1769–1779.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10 (3), 186.

Cabral, J., Krings, M.L., Deco, G., 2017. Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome: models and mechanisms. *NeuroImage* 160, 84–96.

Cai, B., Zille, P., Stephen, J.M., Wilson, T.W., Calhoun, V.D., Wang, Y.P., 2017. Estimation of dynamic sparse connectivity patterns from resting state fMRI. *IEEE Trans. Med. Imaging* 37 (5), 1224–1234.

Calhoun, V.D., Liu, J., Adali, T., 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* 45 (1), S163–S172.

Caporin, M., McAleer, M., 2013. Ten things you should know about the dynamic conditional correlation representation. *Econometrics* 1 (1), 115–126.

Cascio, C., McGlone, F., Folger, S., Tannan, V., Baranek, G., Pelphrey, K.A., Essick, G., 2008. Tactile perception in adults with autism: a multidimensional psychophysical study. *J. Autism Dev. Disord.* 38 (1), 127–137.

Chén, O.Y., Cao, H., Reinen, J.M., Qian, T., Gou, J., Phan, H., De Vos, M., Cannon, T.D., 2019. Resting-state brain information flow predicts cognitive flexibility in humans. *Sci. Rep.* 9 (1), 1–16.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. [arXiv:1412.3555](https://arxiv.org/abs/1412.3555).

Ciric, R., Rosen, A.F., Erus, G., Cieslak, M., Adebeampe, A., Cook, P.A., Bassett, D.S., Davatzikos, C., Wolf, D.H., Satterthwaite, T.D., 2018. Mitigating head motion artifact in functional connectivity MRI. *Nat. Protoc.* 13 (12), 2801–2826.

Cox, R.W., 1996. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29 (3), 162–173.

Cuingnet, R., Glauñès, J.A., Chupin, M., Benali, H., Colliot, O., 2012. Spatial and anatomical regularization of SVM: a general framework for neuroimaging data. *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (3), 682–696.

Dowell, L.R., Mahone, E.M., Mostofsky, S.H., 2009. Associations of postural knowledge and basic motor skill with dyspraxia in autism: implication for abnormalities in distributed connectivity and motor learning. *Neuropsychology* 23 (5), 563.

D'Souza, N.S., Nebel, M.B., Wymbs, N., Mostofsky, S., Venkataraman, A., 2018. A generative-discriminative basis learning framework to predict clinical severity from resting state functional MRI data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 163–171.

Duncan, J., 2005. Frontal lobe function and general intelligence: why it matters.. *Cortex*.

Dziuk, M., Larson, J.G., Apostu, A., Mahone, E., Denckla, M., Mostofsky, S., 2007. Dyspraxia in autism: association with motor, social, and communicative deficits. *Dev. Med. Child Neurol.* 49 (10), 734–739.

D'Souza, N., Nebel, M., Wymbs, N., Mostofsky, S., Venkataraman, A., 2020. A joint network optimization framework to predict clinical severity from resting state functional MRI data. *NeuroImage* 206, 116314.

D'Souza, N.S., Nebel, M.B., Crocetti, D., Wymbs, N., Robinson, J., Mostofsky, S., Venkataraman, A., 2020. A deep-generative hybrid model to integrate multimodal and dynamic connectivity for predicting spectrum-level deficits in autism. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 437–447.

D'Souza, N.S., Nebel, M.B., Wymbs, N., Mostofsky, S., Venkataraman, A., 2019. A coupled manifold optimization framework to jointly model the functional connectomics and behavioral data spaces. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 605–616.

D'Souza, N.S., Nebel, M.B., Wymbs, N., Mostofsky, S., Venkataraman, A., 2019. Integrating neural networks and dictionary learning for multidimensional clinical characterizations from functional connectomics data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 709–717.

Eavani, H., Satterthwaite, T.D., Filipovich, R., Gur, R.E., Gur, R.C., Davatzikos, C., 2015. Identifying sparse connectivity patterns in the brain using resting-state fMRI. *NeuroImage* 105, 286–299.

Engle, R., 2002. Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econ. Stat.* 20 (3), 339–350.

Euston, D.R., Gruber, A.J., McNaughton, B.L., 2012. The role of medial prefrontal cortex in memory and decision making. *Neuron* 76 (6), 1057–1070.

Everson, R., 1998. Orthogonal, but not orthonormal, procrustes problems. *Adv. Comput. Math.* 3 (4).

Feng, C.-M., Gao, Y.-L., Liu, J.-X., Zheng, C.-H., Yu, J., 2017. PCA based on graph Laplacian regularization and p-norm for gene selection and clustering. *IEEE Trans. Nanobiosci.* 16 (4), 257–265.

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8 (9), 700.

Fukushima, M., Betzel, R.F., He, Y., van den Heuvel, M.P., Zuo, X.-N., Sporns, O., 2018. Structure–function relationships during segregated and integrated network states of human brain functional connectivity. *Brain Struct. Funct.* 223 (3), 1091–1106.

Gadgil, S., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Adeli, E., Pohl, K.M., 2020. Spatio-temporal graph convolution for resting-state fMRI analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 528–538.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 315–323.

Goble, D.J., Coxon, J.P., Van Impe, A., Geurts, M., Van Hecke, W., Sunaert, S., Wenderoth, N., Swinnen, S.P., 2012. The neural basis of central proprioceptive processing in older versus younger adults: an important sensory role for right putamen. *Hum. Brain Mapp.* 33 (4), 895–908.

Hahn, K., Myers, N., Prigarin, S., Rodenacker, K., Kurz, A., Förstl, H., Zimmer, C., Wohlschläger, A.M., Sorg, C., 2013. Selectively and progressively disrupted structural connectivity of functional brain networks in Alzheimer's disease-revealed by a novel framework to analyze edge distributions of networks detecting disruptions with strong statistical evidence. *NeuroImage* 81, 96–109.

Havdahl, K.A., Bal, V.H., Huerta, M., Pickles, A., Øyen, A.-S., Stoltenberg, C., Lord, C., Bishop, S.L., 2016. Multidimensional influences on autism symptom measures: implications for use in etiological research. *J. Am. Acad. Child Adolesc. Psychiatry* 55 (12), 1054–1063.

Hearne, L.J., Mattingley, J.B., Cocchi, L., 2016. Functional brain networks related to individual differences in human intelligence at rest. *Sci. Rep.* 6, 32328.

Higgins, I.A., Kundu, S., Guo, Y., 2018. Integrative Bayesian analysis of brain functional networks incorporating anatomical knowledge. *Neuroimage* 181, 263–278.

Honey, C., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.-P., Meuli, R., Hagmann, P., 2009. Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci.* 106 (6), 2035–2040.

Hong, S.-J., Valk, S.L., Di Martino, A., Milham, M.P., Bernhardt, B.C., 2018. Multidimensional neuroanatomical subtyping of autism spectrum disorder. *Cereb. Cortex* 28 (10), 3578–3588.

Insel, T.R., 2014. The nihm research domain criteria (RDOC) project: precision medicine for psychiatry. *Am. J. Psychiatry* 171 (4), 395–397.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. *Neuroimage* 62 (2), 782–790.

Kaiser, M.D., Hudac, C.M., Shultz, S., Lee, S.M., Cheung, C., Berken, A.M., Deen, B., Pitskel, N.B., Sugrue, D.R., Voos, A.C., et al., 2010. Neural signatures of autism. *Proc. Natl. Acad. Sci.* 201010412.

Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. Brainnetcnn: Convolutional neural networks for brain networks: towards predicting neurodevelopment. *NeuroImage* 146, 1038–1049.

Kiar, G., Roncal, W.G., Mbembe, D., Bridgeford, E., Burns, R., Vogelstein, J., 2016. ndmg: neurodata's mri graphs pipeline. Zenodo.

Kingma, D. P., Ba, J. L., 2015. Adam: a method for stochastic optimization.

Koshino, H., Carpenter, P.A., Minshew, N.J., Cherkassky, V.L., Keller, T.A., Just, M.A., 2005. Functional connectivity in an fMRI working memory task in high-functioning autism. *Neuroimage* 24 (3), 810–821.

Lindquist, M., 2016. Dynamic connectivity: pitfalls and promises.

Lindquist, M.A., Xu, Y., Nebel, M.B., Caffo, B.S., 2014. Evaluating dynamic bivariate correlations in resting-state fMRI: a comparison study and a new approach. *NeuroImage* 101, 531–546.

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L., DiLavore, P.C., Pickles, A., Rutter, M., 2000. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* 30 (3), 205–223.

Lynch, C.J., Uddin, L.Q., Supekar, K., Khouzam, A., Phillips, J., Menon, V., 2013. Default mode network in childhood autism: posteromedial cortex heterogeneity and relationship with social deficits. *Biol. Psychiatry* 74 (3), 212–219.

Manton, J.H., Mahony, R., Hua, Y., 2003. The geometry of weighted low-rank approximations. *IEEE Trans. Signal Process.* 51 (2), 500–514.

Menon, V., 2011. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cognit. Sci.* 15 (10), 483–506.

Mostofsky, S.H., Dubey, P., Jerath, V.K., Jansiewicz, E.M., Goldberg, M.C., Denckla, M.B., 2006. Developmental dyspraxia is not limited to imitation in children with autism spectrum disorders. *J. Int. Neuropsychol. Soc.* 12 (3), 314–326.

Muschelli, J., Nebel, M.B., Caffo, B.S., Barber, A.D., Pekar, J.J., Mostofsky, S.H., 2014. Reduction of motion-related artifacts in resting state fMRI using acompcor. *Neuroimage* 96, 22–35.

Nebel, M.B., Eloyan, A., Nettles, C.A., Sweeney, K.L., Ament, K., Ward, R.E., Choe, A.S., Barber, A.D., Pekar, J.J., Mostofsky, S.H., 2016. Intrinsic visual-motor synchrony correlates with social deficits in autism. *Biol. Psychiatry* 79 (8), 633–641.

Niznikiewicz, M.A., Kubicki, M., Shenton, M.E., 2003. Recent structural and functional imaging findings in schizophrenia. *Curr. Opin. Psychiatry* 16 (2), 123–147.

Nocedal, J., Wright, S., 2006. Numerical Optimization. Springer Science & Business Media.

Pang, J., Cheung, G., 2017. Graph Laplacian regularization for image denoising: analysis in the continuous domain. *IEEE Trans. Image Process.* 26 (4), 1770–1785.

Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D., 2018. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med. Image Anal.* 48, 117–130.

Park, C.-h., Kim, S.Y., Kim, Y.-H., Kim, K., 2008. Comparison of the small-world topology between anatomical and functional connectivity in the human brain. *Physica A* 387 (23), 5958–5962.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical Parametric Mapping: the Analysis of Functional Brain Images. Elsevier.

Pouw, L.B., Rieffe, C., Stockmann, L., Gadow, K.D., 2013. The link between emotion regulation, social functioning, and depression in boys with ASD. *Res. Autism Spect. Disord.* 7 (4), 549–556.

Price, T., Wee, C.-Y., Gao, W., Shen, D., 2014. Multiple-network classification of childhood autism using functional connectivity dynamics. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 177–184.

Propper, R.E., O'Donnell, L.J., Whalen, S., Tie, Y., Norton, I.H., Suarez, R.O., Zollei, L., Radmanesh, A., Golby, A.J., 2010. A combined fMRI and DTI examination of functional language lateralization and arcuate fasciculus structure: effects of degree versus direction of hand preference. *Brain Cognit.* 73 (2), 85–92.

Rabany, L., Brocke, S., Calhoun, V.D., Pittman, B., Corbera, S., Wexler, B.E., Bell, M.D., Pelphrey, K., Pearson, G.D., Assaf, M., 2019. Dynamic functional connectivity in schizophrenia and autism spectrum disorder: convergence, divergence and classification. *NeuroImage* 24, 101966.

Raichle, M.E., 2015. The brain's default mode network. *Annu. Rev. Neurosci.* 38, 433–447.

Rashid, B., Damaraju, E., Pearson, G.D., Calhoun, V.D., 2014. Dynamic connectivity states estimated from resting fMRI identify differences among schizophrenia, bipolar disorder, and healthy control subjects. *Front. Hum. Neurosci.* 8, 897.

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52 (3), 1059–1069.

Rudie, J.D., Brown, J., Beck-Pancer, D., Hernandez, L., Dennis, E., Thompson, P., Bookheimer, S., Dapretto, M., 2013. Altered functional and structural brain network organization in autism. *NeuroImage* 2, 79–94.

Savva, A.D., Mitsis, G.D., Matsopoulos, G.K., 2019. Assessment of dynamic functional connectivity in resting-state fMRI using the sliding window technique. *Brain Behav.* 9 (4), e01255.

Schnabel, R.B., Toint, P.L., 1983. Forcing sparsity by projecting with respect to a non-diagonally weighted frobenius norm. *Math. Program.* 25 (1), 125–129.

Sestieri, C., Corbetta, M., Romani, G.L., Shulman, G.L., 2011. Episodic memory retrieval, parietal cortex, and the default mode network: functional and topographic analyses. *J. Neurosci.* 31 (12), 4407–4420.

Skudlarski, P., Jagannathan, K., Calhoun, V.D., Hampson, M., Skudlarska, B.A., Pearson, G., 2008. Measuring brain connectivity: diffusion tensor imaging validates resting state temporal correlations. *Neuroimage* 43 (3), 554–561.

Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., et al., 2013. Resting-state fMRI in the human connectome project. *Neuroimage* 80, 144–168.

Spitzer, R.L., Williams, J.B., 1980. Diagnostic and statistical manual of mental disorders. American Psychiatric Association. Citeseer.

Sporns, O., Chialvo, D.R., Kaiser, M., Hilgetag, C.C., 2004. Organization, development and function of complex brain networks. *Trends Cognit. Sci.* 8 (9), 418–425.

Sridharan, D., Levitin, D.J., Menon, V., 2008. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc. Natl. Acad. Sci.* 105 (34), 12569–12574.

Sui, J., He, H., Yu, Q., Rogers, J., Pearson, G., Mayer, A.R., Bustillo, J., Canive, J., Calhoun, V.D., et al., 2013. Combination of resting state fMRI, DTI, and SMRI data to discriminate schizophrenia by n-way mcca + jica. *Front. Hum. Neurosci.* 7, 235.

Sun, Y., Yin, Q., Fang, R., Yan, X., Wang, Y., Bezerianos, A., Tang, H., Miao, F., Sun, J., 2014. Disrupted functional brain connectivity and its association to structural connectivity in amnestic mild cognitive impairment and Alzheimer's disease. *PLoS One* 9 (5).

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM: using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage* 15 (1), 273–289.

Uddin, L.Q., Yeo, B.T., Spreng, R.N., 2019. Towards a universal taxonomy of macro-scale functional human brain networks. *Brain Topogr.* 1–17.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., et al., 2013. The wu-minn human connectome project: an overview. *Neuroimage* 80, 62–79.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., et al., 2012. The human connectome project: a data acquisition perspective. *Neuroimage* 62 (4), 2222–2231.

Venkataraman, A., Duncan, J.S., Yang, D.Y.-J., Pelphrey, K.A., 2015. An unbiased bayesian approach to functional connectomics implicates social-communication networks in autism. *NeuroImage: Clinical* 8, 356–366.

Venkataraman, A., Kubicki, M., Golland, P., 2012. From brain connectivity models to identifying foci of a neurological disorder. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 715–722.

Venkataraman, A., Kubicki, M., Golland, P., 2013. From connectivity models to region labels: identifying foci of a neurological disorder. *IEEE Trans. Med. Imaging* 32 (11), 2078–2098.

Venkataraman, A., Rathi, Y., Kubicki, M., Westin, C.-F., Golland, P., 2011. Joint modeling of anatomical and functional connectivity for population studies. *IEEE Trans. Med. Imaging* 31 (2), 164–182.

Venkataraman, A., Wymbs, N., Nebel, M.B., Mostofsky, S., 2017. A unified bayesian approach to extract network-based functional differences from a heterogeneous patient cohort. In: International Workshop on Connectomics in Neuroimaging. Springer, pp. 60–69.

Venkataraman, A., Yang, D.Y.-J., Pelphrey, K.A., Duncan, J.S., 2016. Bayesian community detection in the space of group-level functional differences. *IEEE Trans. Med. Imaging* 35 (8), 1866–1882.

Vissers, M.E., Cohen, M.X., Geurts, H.M., 2012. Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links. *Neurosci. Biobehav. Rev.* 36 (1), 604–625.

Wang, F., Kalmar, J.H., He, Y., Jackowski, M., Chepenik, L.G., Edmiston, E.E., Tie, K., Gong, G., Shah, M.P., Jones, M., et al., 2009. Functional and structural connectivity between the perigenual anterior cingulate and amygdala in bipolar disorder. *Biol. Psychiatry* 66 (5), 516–521.

Wang, Q., Su, T.-P., Zhou, Y., Chou, K.-H., Chen, I.-Y., Jiang, T., Lin, C.-P., 2012. Anatomical insights into disrupted small-world networks in schizophrenia. *Neuroimage* 59 (2), 1085–1093.

Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2012. Identification of MCI individuals using structural and functional connectivity networks. *Neuroimage* 59 (3), 2045–2056.

Weyandt, L., Swentosky, A., Gudmundsdottir, B.G., 2013. Neuroimaging and ADHD: fMRI, pet, DTI findings, and methodological limitations. *Dev. Neuropsychol.* 38 (4), 211–225.

Whitwell, J.L., Avula, R., Master, A., Vemuri, P., Senjem, M.L., Jones, D.T., Jack Jr, C.R., Josephs, K.A., 2011. Disrupted thalamocortical connectivity in PSP: a resting-state fMRI, DTI, and VBM study. *Parkinsonism Relat. Disord.* 17 (8), 599–605.

Williams, D.L., Goldstein, G., Minshew, N.J., 2006. The profile of memory function in children with autism. *Neuropsychology* 20 (1), 21.

Zimmermann, J., Griffiths, J.D., McIntosh, A.R., 2018. Unique mapping of structural and functional connectivity on cognition. *J. Neurosci.* 38 (45), 9658–9667.